UNIVERSIDADE ESTADUAL DE CAMPINAS

INSTITUTO DE BIOLOGIA

FELIPE ALONSO MARTINS

APLICAÇÃO DE FERRAMENTAS DE BIOLOGIA DE SISTEMAS EM LEVEDURA INDUSTRIAL PARA PRODUÇÃO DE BIOETANOL DE SEGUNDA GERAÇÃO

CAMPINAS

2016

FELIPE ALONSO MARTINS

APLICAÇÃO DE FERRAMENTAS DE BIOLOGIA DE SISTEMAS EM LEVEDURA INDUSTRIAL PARA PRODUÇÃO DE BIOETANOL DE SEGUNDA GERAÇÃO

Dissertação apresentada ao Instituto de Biologia da Universidade Estadual de Campinas como parte dos requisitos exigidos para a obtenção do Título de Mestre em Genética e Biologia Molecular, na área de Bioinformática.

ESTE ARQUIVO DIGITAL CORRESPONDE À VERSÃO FINAL DA DISSERTAÇÃO DEFENDIDA PELO ALUNO FELIPE ALONSO MARTINS E ORIENTADA PELO PROF. DR. GONÇALO AMARANTE GUIMARÃES PEREIRA.

Orientador: PROF. DR. GONÇALO AMARANTE GUIMARÃES PEREIRA

CAMPINAS

2016

Agência(s) de fomento e nº(s) de processo(s): CAPES

Ficha catalográfica
Universidade Estadual de Campinas
Biblioteca do Instituto de Biologia
Mara Janaina de Oliveira - CRB 8/6972

M366a

Martins, Felipe Alonso, 1989-
Aplicação de ferramentas de biologia de sistemas em levedura industrial para produção de bioetanol de segunda geração / Felipe Alonso Martins. – Campinas, SP : [s.n.], 2016.

Orientador: Gonçalo Amarante Guimarães Pereira.
Dissertação (mestrado) – Universidade Estadual de Campinas, Instituto de Biologia.

1. *Saccharomyces cerevisiae*. 2. Xilose. 3. Etanol 2G. 4. Bioinformática. 5. Biologia de sistemas. I. Pereira, Gonçalo Amarante Guimarães,1964-. II. Universidade Estadual de Campinas. Instituto de Biologia. III. Título.

Informações para Biblioteca Digital

Título em outro idioma: Application of systems biology tools in industrial yeast for second-generation bioethanol production
Palavras-chave em inglês:
Saccharomyces cerevisiae
Xylose
2G ethanol
Bioinformatics
Systems biology
Área de concentração: Bioinformática
Titulação: Mestre em Genética e Biologia Molecular
Banca examinadora:
Gonçalo Amarante Guimarães Pereira [Orientador]
Renato Vicentini dos Santos
Diego Mauricio Riaño-Pachón
Data de defesa: 23-02-2016
Programa de Pós-Graduação: Genética e Biologia Molecular

Campinas, 23 de fevereiro de 2016.

COMISSÃO EXAMINADORA

Prof. Dr. Gonçalo Amarante Guimarães Pereira
(Orientador)

Prof. Dr. Renato Vicentini dos Santos

Prof. Dr. Diego Mauricio Riaño-Pachón

Os membros da Comissão Examinadora acima assinaram a Ata de Defesa, que se encontra no processo de vida acadêmica do aluno.

Dedico este trabalho a todas as pessoas que de alguma forma ajudaram a moldar minha forma de pensar e me tornaram o que sou hoje. Em especial meus pais Maura e Edison, meus irmãos Guilherme e Henrique, minha namorada Mizuho, meus tios, primos e avós, meus amigos da Casa Amarela, meus amigos de infância e adolescência e todos meus professores.

AGRADECIMENTOS

Agradeço ao meu orientador Prof. Dr. Gonçalo pela confiança, apoio e entusiasmo desde meus tempos de bolsista de graduação.

Ao meu coorientador Dr. Marcelo, pela imensa ajuda, dicas e inspiração durante todo o trabalho.

A todas as pessoas que passaram pelo Laboratório de Genômica e Expressão, pelo apoio e amizade.

Ao professor Dr. Paulo de Oliveira e todos os outros membros da Bioinformática do LNBio, pelo acolhimento e imensa ajuda durante minha pesquisa lá.

Aos meus pais, por todo amor e apoio durante meus anos de Unicamp.

Ao meu irmão Guilherme, pela imensa ajuda no dia-a-dia, organização e paciência.

Ao meu irmão Henrique, por não me deixar perder meu lado criança.

À minha namorada Mizuho, por todo amor, carinho e companheirismo, e por sempre me acalmar nos momentos mais difíceis.

Aos moradores da Casa Amarela, que me acompanham desde o início de Unicamp, pela amizade e diversão

À Fundação Capes por financiar este trabalho.

RESUMO

O Brasil é um dos líderes mundiais na produção de etanol, porém o país já enfrenta uma grande limitação imposta pela tecnologia de primeira geração. Assim, novas alternativas vêm sendo propostas, com destaque para a tecnologia de segunda geração, que consiste em utilizar os resíduos lignocelulósicos da cana-de-açúcar para a produção de etanol. Um dos maiores desafios dessa nova tecnologia é o desenvolvimento de uma levedura industrial capaz de produzir etanol a partir da xilose existente no material lignocelulósico. Apesar de estudos mostrarem resultados promissores na obtenção de etanol com micro-organismos naturalmente capazes de fermentar tanto hexoses (glicose) quanto pentoses (xilose), nenhum deles possui a mesma capacidade fermentativa, tolerância a etanol e robustez da *S. cerevisiae*. Logo, a modificação genética desta espécie para utilizar as vias metabólicas de assimilação de xiloses surge como um dos maiores desafios atuais para alavancar a produção de bioetanol no Brasil.

A fim de tornar possível a fermentação de xilose por *S. cerevisiae*, é necessária a manipulação do seu genoma para a inserção de alguma via que permita a conversão de xilose em xilulose, sendo que são conhecidas duas vias capazes de realizar essa conversão: a via oxi-redutiva, composta pelas enzimas xilose redutase (XR) e xilitol desidrogenase (XDH), e a via da isomerização, composta pela enzima xilose isomerase (XI). Ambas apresentam limitações quando expressas em *S. cerevisiae*, a via oxi-redutiva apresenta um desequilíbrio nos cofatores das enzimas, o que provoca acúmulo de xilitol e limitação na produção de etanol. A via da isomerização não apresenta tal limitação, porém a enzima da xilose isomerase não possui atividade catalítica quando expressada em *S. cerevisiae*, provavelmente devido a incorreto enovelamento dessas proteínas.

Neste trabalho foram utilizadas ferramentas de biologia de sistemas para simular modelos metabólicos em escala genômica de fermentação de xilose por *S. cerevisiae*, a fim de estudar o desequilíbrio de cofatores da via oxi-redutiva e identificar possíveis alvos de engenharia genética. As simulações foram realizadas utilizando a abordagem de análise de balanço de fluxo (FBA) que prediz, através de programação dinâmica, o máximo valor possível de uma determinada função objetivo (que normalmente representa a produção de biomassa) de um sistema linear com equações e restrições representando o fluxo de metabólitos em uma enzima. Também foram realizados trabalhos de mineração de sequências em

bancos de dados públicos, na tentativa de encontrar xilose isomerases com potencial de terem uma atividade catalítica em *S. cerevisiae,* além de trabalhos de modelagem de proteínas por homologia, utilizando proteínas funcionais e não funcionais, a fim de identificar características estruturais que determinam sua funcionalidade em *S. cerevisiae*.

ABSTRACT

Brazil is a world leader in ethanol production, but the country is already facing a major limitation imposed by first-generation ethanol production technology. Thus, new alternatives are being proposed, notably the second-generation technology, which consists in use sugarcane lignocellulosic residues for ethanol production. One of the major challenges to this new technology is the development of a yeast capable of producing ethanol from the xylose present in the lignocellulosic material. Although studies showing promising results with microorganisms naturally capable of ferment both hexoses (glucose) and pentoses (xylose) sugars to ethanol, none of them has the same fermentative capacity, ethanol tolerance and robustness than *S. cerevisiae*. Therefore, the genetic modification of this species to insert metabolic pathways to consume xylose appears as one of the biggest current challenges to boost the bioethanol production in Brazil.

In order to make possible the xylose fermentation by *S. cerevisiae*, the insertion of some pathway that converts xylose to xylulose is necessary, being that two pathways are known as capable of performing this conversion: the oxido-reductase pathway, formed by the enzymes xylose reductase (XR) and xylitol dehydrogenase (XDH), and the isomerase pathway, formed by the enzyme xylose isomerase (XI). Both pathways present limitations when expressed in *S. cerevisiae*, the oxido-reductase pathway has a problem related to cofactor imbalance of its enzymes, what causes xylitol accumulation and limitation in ethanol production. The isomerase pathway does not show this limitation, but even so, the enzyme xylose isomerase does not present catalytic activity when expressed in *S. cerevisiae*, probably due to incorrect folding of this protein.

In this work, systems biology tools were used to simulate xylose fermentation in genome-scale metabolic models of *S. cerevisiae*, in order to study the cofactors imbalance of the oxido-reductase pathway and identify possible genetic engineering targets. The simulations were performed with flux balance analysis (FBA) approach, that predicts, with dynamic programming, the maximum value possible of a defined objective function (which usually represents the biomass production) of a linear system with equations and restrictions representing the metabolic flux through an enzyme. In addition, data mining in public databases was done in attempt to find xyloses isomerases with potential to have catalytic activity in *S. cerevisiae*, besides

works of protein homology modeling based on functional and non-functional proteins in order to identify structural features that define its functionality in *S. cerevisiae*.

SUMÁRIO

Capítulo 1

INTRODUÇÃO

1.1 ETANOL

Atualmente, o etanol é o biocombustível mais consumido no mundo, sendo que o Brasil se destaca na produção ao lado dos Estados Unidos – produção em 2011 estimada em 21 e 52 bilhões de litros, respectivamente (RENEWABLE FUELS ASSOCIATION, 2012). No Brasil, a produção do chamado etanol de primeira geração ocorre diretamente pela fermentação da sacarose presente no caldo da cana-de-açúcar realizada por cepas da levedura *Saccharomyces cerevisiae*, sem necessidade de nenhum pré-tratamento.

Para uma melhor eficiência na fermentação, algumas cepas foram selecionadas naturalmente, estando disponíveis desde os anos 80, porém, devido ao ambiente industrial hostil, frequentemente elas acabam não resistindo e são substituídas durante a fermentação por cepas selvagens ineficientes que residem na cana-de-açúcar. Assim, uma nova estratégia foi adotada nos anos 90, e cepas passaram a ser selecionadas entre as selvagens que apresentavam uma alta eficiência na fermentação aliada a uma grande resistência frente ao sistema industrial. Essas linhagens, comumente denominadas “industriais”, vêm sendo largamente adotadas no ambiente industrial como inóculos iniciais para a produção de etanol. Como exemplo de grande sucesso podemos citar a cepa PE-2, usada por cerca de 30% das destilarias brasileiras, sendo responsável por aproximadamente 10% do suprimento mundial de etanol (ARGUESO et al., 2009).

A produção brasileira se mostra altamente competitiva economicamente quando comparada à produção através de milho e beterraba, feita nos Estados Unidos e Europa, respectivamente. Na produção a partir do milho, é necessário o uso de enzimas para transformar o amido presente no grão em açúcares fermentescíveis, aumentando o custo de produção. Já no caso da beterraba, o baixo teor de sacarose e a baixa produtividade, atrapalham a competitividade do produto (GOLDEMBERG, 2008).

Mesmo com as vantagens da cana-de-açúcar, o Brasil ainda não é autossuficiente na produção de etanol e o preço do combustível no mercado

apresenta grandes oscilações durante o ano, principalmente devido ao período de entressafras. Somando a isso o fato de o país apresentar um grande mercado potencial para consumo de etanol, fica evidente a necessidade de aumentar sua produção, não só através da expansão das fronteiras agrícolas, mas também com a criação de alternativas economicamente viáveis, que permitam aumentar o rendimento sem um aumento significativo nos custos de produção. Dentre as alternativas, estão o uso de variedades vegetais geneticamente modificadas, utilização de resíduos da agricultura como matéria-prima (etanol de segunda geração) e melhoramento da eficiência da fermentação das linhagens de leveduras.

1.2 ETANOL DE SEGUNDA GERAÇÃO

O etanol de segunda geração, ou etanol celulósico, é considerado o maior produto da biotecnologia industrial do mundo, sendo que seu desenvolvimento apresenta resultados muito promissores. A tecnologia de segunda geração consiste na obtenção de etanol a partir de resíduos lignocelulósicos (celulose, hemicelulose e lignina, cuja associação forma a parede celular vegetal) gerados a partir da produção tradicional de etanol (OTERO; NIELSEN, 2010). O processo de produção do etanol de segunda geração está descrito na figura 1.1.

A celulose, um polímero de cadeias longas e fortemente ligadas, composto exclusivamente de glicose, é o componente mais abundante da parede celular; já a hemicelulose é uma mistura de polímeros de menor tamanho e formados tanto por hexoses, como glicose e galactose, quanto por pentoses, como xilose e arabinose, sendo que a sua composição varia bastante entre espécies vegetais; por fim, a lignina é um complexo polímero hidrofóbico muito resistente à degradação química ou biológica (RUDOLF et al., 2009).

Um dos desafios da tecnologia de segunda geração reside na dificuldade em se obter os açúcares para fermentação devido às complexas e rígidas estruturas da parede celular, que atrapalham a hidrólise do material lignocelulósico. Para tanto, são realizados pré-tratamentos que visam a separação das cadeias da parede e aumento da porosidade do bagaço, facilitando a hidrólise da celulose e hemicelulose em monossacarídeos, permitindo a fermentação alcoólica. As hexoses são facilmente fermentadas pela *S. cerevisiae*, em um processo bem explorado e

estabelecido, porém, a levedura é incapaz de metabolizar as pentoses, que representam de 15% a 45% do material lignocelulósico (KOOTSTRA et al., 2009).

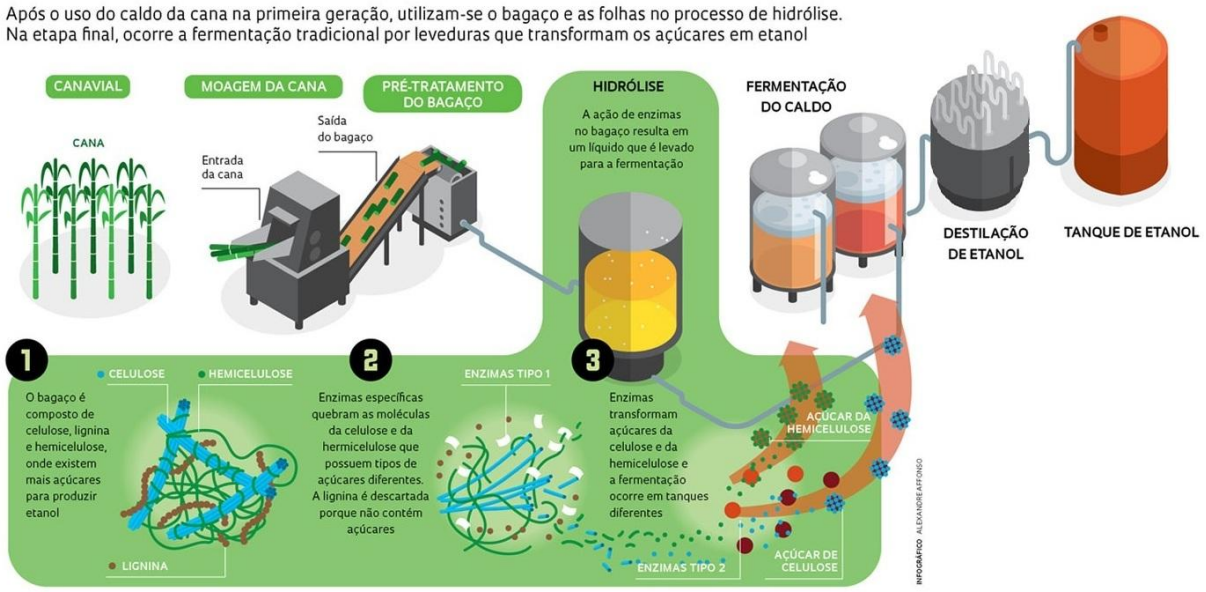


Figura 1.1 (OLIVEIRA, 2012): Processo de produção do etanol de segunda geração

Dessa forma um grande desafio da tecnologia de segunda geração reside na obtenção de leveduras industriais capazes de fermentar tanto pentoses como hexoses, o que permitiria aumentar consideravelmente a produtividade e reduzir os custos na produção do bioetanol. Apesar de estudos mostrarem resultados promissores na obtenção de etanol com micro-organismos naturalmente capazes de fermentar ambos os tipos de açúcares como *Pichia stipitis* (AGBOGBO et al., 2006), *Candida shehatae* (CHANDEL et al., 2007), *Pachysolen tannophilus* (CHENG et al., 2008), *Kluyveromyces marxianus* (MARGARITIS; BAJPAI, 1982), entre outros, nenhum possui a mesma capacidade fermentativa, tolerância a etanol e robustez que *S. cerevisiae* (BALAT, 2011). Logo, a modificação genética desta espécie para utilizar as vias metabólicas de assimilação de pentoses surge como um dos maiores desafios atuais a fim de se obter uma produção satisfatória de bioetanol (GÍRIO et al., 2010).

1.3 LEVEDURA INDUSTRIAL BRASILEIRA – PEDRA II

Apesar da levedura *S. cerevisiae* ter seu uso amplamente difundido na indústria brasileira, suas cepas isoladas permaneciam completamente desconhecidas do ponto de vista genético e molecular até pouco tempo atrás. Essa

falta de informações básicas sobre a biologia dessas cepas representa uma barreira para o eventual melhoramento de suas habilidades naturais. Recentemente, o Laboratório de Genômica e Expressão da Unicamp (LGE) e colaboradores (ARGUESO et al., 2009) sequenciou o genoma completo de um esporo derivado de Pedra II (PE-2). Este trabalho teve importante impacto científico para o setor, uma vez que ineditamente uma linhagem brasileira utilizada em larga escala na produção de etanol teve sua estrutura genética descrita em detalhes e as análises do genoma podem explicar parcialmente o ótimo desempenho fermentativo da linhagem PE-2 observado nas dornas. Além disso, estes dados forneceram conhecimentos para planejar estratégias de manipulação genética que não interfiram com suas características desejáveis permitindo o uso dessa cepa como plataformas biológicas versáteis e otimizadas para aplicações diversas em indústrias de biotecnologia.

Com o intuito de complementar as informações geradas pelo genoma de PE-2 foram desenvolvidos diversos trabalhos de transcriptômica utilizando a metodologia de RNA-seq. Estes projetos produziram um atlas de informações sobre expressão gênica global de PE-2 produzindo etanol sob diversas condições controladas de laboratório e também em escala industrial dentro das usinas (CARVALHO-NETTO et al., 2015).

1.4 VIAS METABÓLICAS PARA CONSUMO DE XILOSE

A fermentação de xilose por *S. cerevisiae* foi obtida pela primeira vez com a expressão heteróloga da via oxi-redutiva presente na levedura *Pichia stipitis* em conjunto com a superexpressão de xiluloquinase (XK) endógena (KARHUMA et al., 2007). Esta via, representada na figura 1.2, utiliza xilose redutase (XR) e xilitol desidrogenase (XDH) para converter, respectivamente, xilose em xilitol e xilitol em xilulose que, após ser fosforilada pela xiluloquinase, é metabolizada através da via das pentoses-fosfato (PPP) (BALAT, 2011). Porém, as diferentes preferências de cofatores entre XR (NADPH) e XDH (NAD+) limitam o fluxo de xilose para xilulose, gerando a secreção indesejada do intermediário xilitol (KARHUMA et al., 2007).

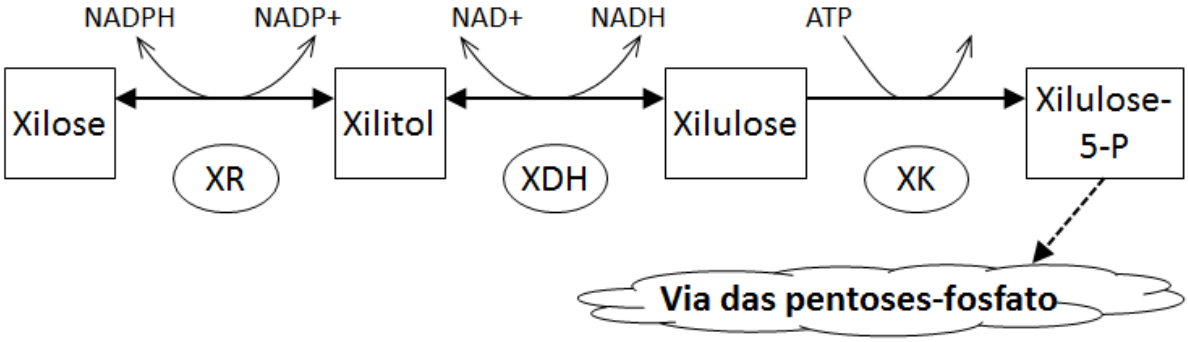


Figura 1.2: Via oxi-redutiva para consumo de xilose. Xilose redutase (XR) converte xilose em xilitol, xilitol desidrogenase (XDH) converte xilitol em xilulose, que é fosforilada pela xiluloquinase e metabolizada através da via das pentoses-fosfato (PPP).

Algumas soluções já foram propostas para a correção do balanço redox da via oxi-redutiva. Um dos exemplos é um estudo que incluiu a via da Fosfoquetolase (figura 1.3) para catabolismo de pentoses na *S. cerevisiae* através de engenharia metabólica (SONDEREGGER et al, 2004). Considerando que o maior obstáculo para o funcionamento da via oxi-redutiva é a capacidade limitada da levedura na reoxidação de NADH, o estudo se propõe a solucionar o problema canalizando os fluxos através de uma via de fosfoquetolase recombinante, formada, além da fosfoquetolase, pelas enzimas fosfotransacetilase e acetaldeído desidrogenase. Como consequência o rendimento de etanol foi aumentado em 25% devido a uma redução da formação do subproduto xilitol. O fluxo sobre a via da fosfoquetolase foi cerca de 30% do que seria necessário para eliminar completamente a formação de glicerol e xilitol.

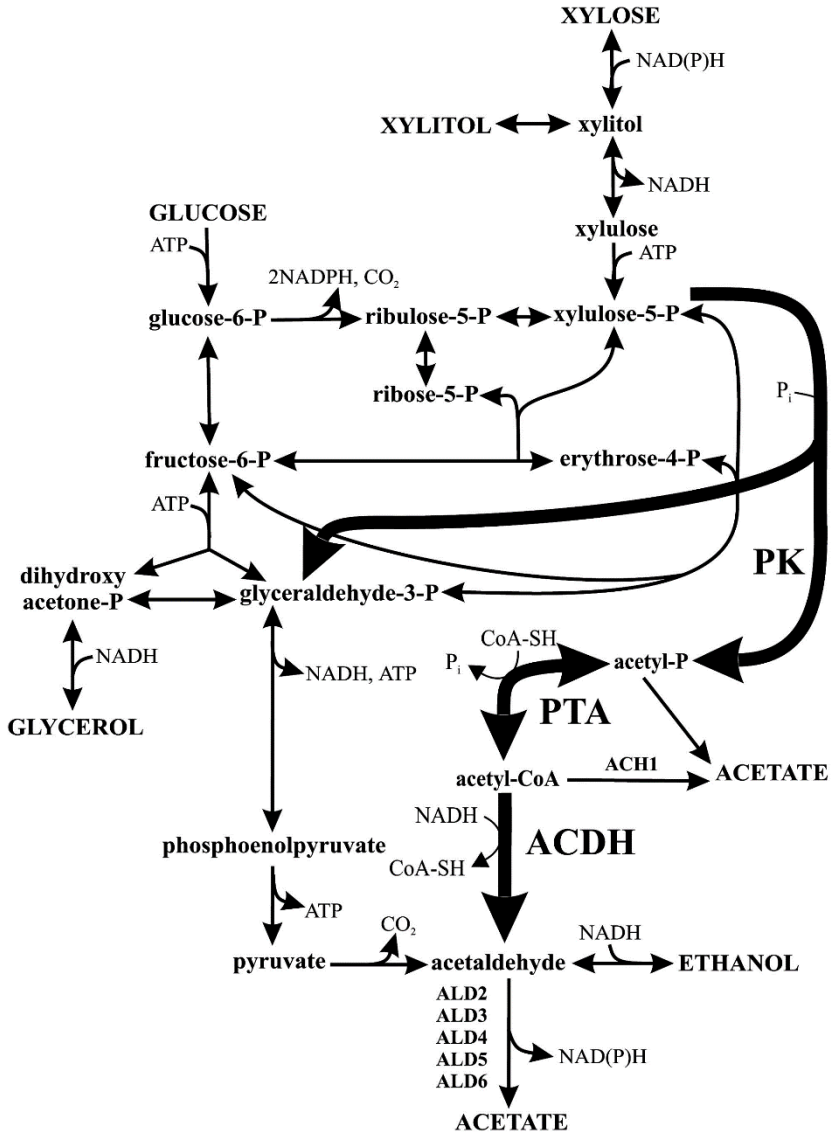


Figura 1.3 (SONDEREGGER et al, 2004): Trecho da rede metabólica da *S. cerevisiae* com expressão das vias oxi-redutiva e da fosfoquetolase. Abreviações: PK, fosfoquetolase; PTA, fosfotransacetilase; ACDH, acetaldeído desidrogenase.

A expressão heteróloga de xilose isomerase (XI) é outra opção para conversão de xilose em xilulose. Esta via está amplamente presente em diversas bactérias e em somente poucos fungos – provavelmente, oriundos de transferência horizontal de bactéria (GÁRDONYI; HAHN-HÄGERDAL, 2003) – sendo que os melhores resultados em *S. cerevisiae* foram obtidos utilizando o gene presente no fungo *Piromyces sp*. Porém, apenas a expressão de XI (em conjunto com a superexpressão endógena já citada de XK) não resultou em um crescimento satisfatório em xilose, sendo necessárias várias adaptações (KUYPER et al., 2004) ou engenharia genética (KUYPER et al., 2005) que proporcionaram uma velocidade

de consumo de xilose maior, obtendo crescimento e fermentação de forma satisfatória (KARHUMAA et al., 2007). A figura 1.4 apresenta o funcionamento da via.

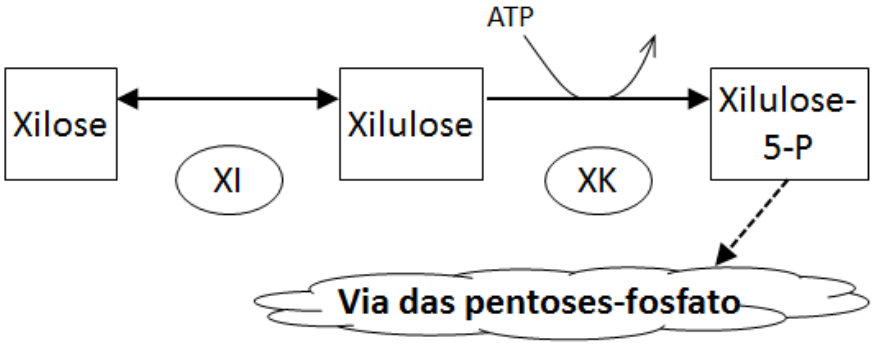


Figura 1.4: Via da isomerização de xilose. Xilose isomerase (XI) converte xilose diretamente em xilulose, que é fosforilada pela xiluloquinase e metabolizada através da via das pentoses-fosfato (PPP).

No caso da via da isomerização, uma questão muito interessante é o fato da grande maioria das XI bacterianas não possuir nenhuma atividade catalítica quando expressa em *Saccharomyces cerevisiae*. Isso se deve muito provavelmente ao enovelamento incorreto das proteínas de origem bacteriana dentro da célula eucariótica da levedura (GÁRDONYI, HAHN-HÄGERDAL, 2003).

1.5 BIOINFORMÁTICA E BIOLOGIA DE SISTEMAS

Visto a grande quantidade de dados obtidos através do sequenciamento de elementos genômicos e transcriptômicos e com os avanços das ferramentas de simulação computacional, a bioinformática tem cada vez mais um papel fundamental na biologia, possibilitando o estudo *in silico* dos processos metabólicos de uma forma mais ampla, indicando gargalos e também novas modificações genéticas que permitam otimizar fluxos metabólicos de interesse.

Nesse contexto, a biologia de sistemas (também conhecida como biologia sistêmica), que consiste na análise quantitativa de sistemas biológicos, principalmente através de modelos matemáticos preditivos, permite a análise e integração de todos os conjuntos de dados em escala genômica a fim de obter uma descrição quantitativa do fenótipo do sistema (OTERO; NIELSEN, 2010). Através da

biologia de sistemas a rede metabólica do organismo pode ser modelada *in silico*, e assim ser feita a predição dos fenótipos causados (alteração na produção de metabólitos) por alterações genéticas no organismo, realizando-se testes *in vivo* apenas para validar os resultados de maior interesse.

Um modelo metabólico em escala genômica é basicamente uma representação estequiométrica de todas as possíveis reações metabólicas de uma célula. A partir deste modelo, tendo em mente a reversibilidade das reações e assumindo um estado estacionário dos metabólitos internos, é possível determinar os possíveis caminhos que levam do substrato aos metabólitos finais inferindo os fluxos em cada etapa. Muitos algoritmos vêm sendo desenvolvidos com base nessa metodologia para auxiliar na identificação de genes alvo de engenharia genética que possibilitem o aumento do fluxo metabólico na direção de algum metabólito específico (LIU et al., 2010).

Dentre os mais utilizados, temos a análise de balanço de fluxo (FBA – *flux balance analysis*) que calcula os valores de todos os fluxos do modelo a partir de uma matriz estequiométrica, sendo assim capaz de predizer as taxas de substratos consumidas, taxas de subprodutos produzidos e taxas de crescimento. Um outro importante algoritmo é a minimização de ajuste metabólico (MOMA) que prediz as taxas de consumo e produção de metabólitos para um mutante (criado a partir de eventos de deleção ou inserção de genes) utilizando a distribuição de fluxos calculado a partir de FBA do organismo selvagem. A figura 1.5 apresenta a evolução dos algoritmos e ferramentas que realizam FBA.

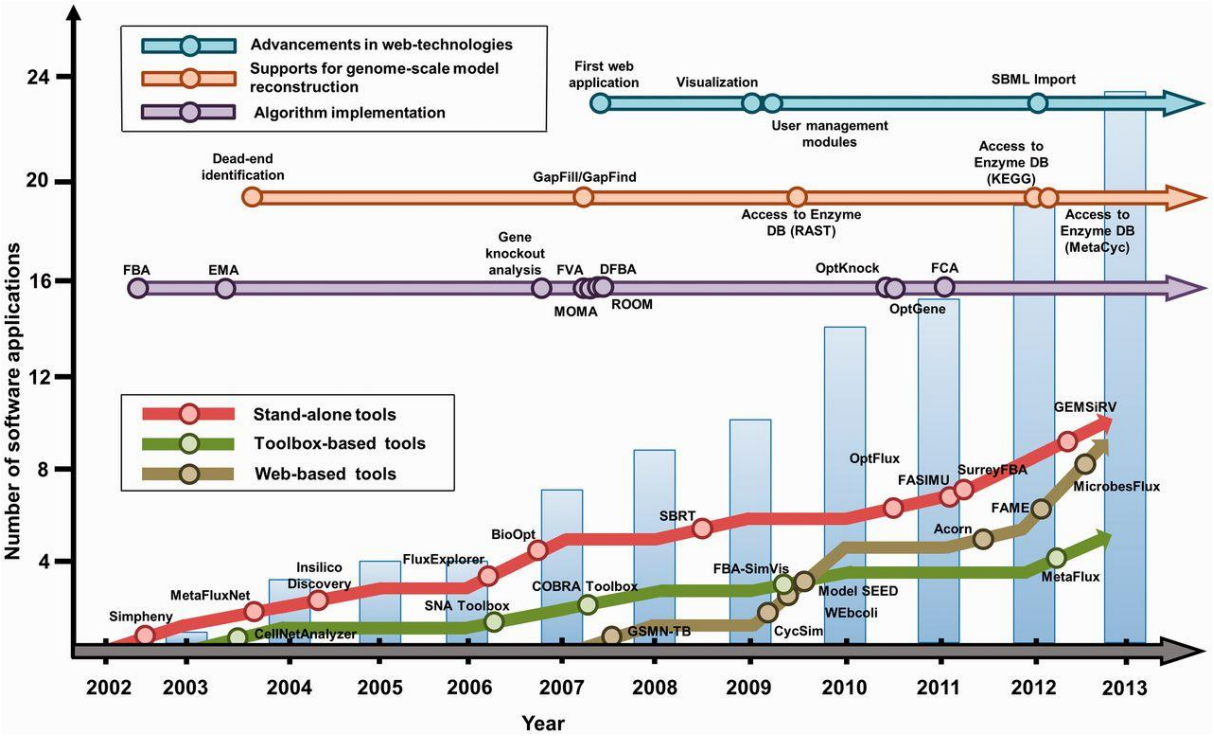


Figura 1.5 (LAKSHMANAN et al., 2012): Evolução dos algoritmos e programas que implementam FBA.

Alguns modelos metabólicos já vêm sendo empregados visando a otimização da produção de etanol de segunda geração utilizando como base a levedura *S. cerevisiae* (OTERO; NIELSEN, 2010). Por exemplo, um modelo metabólico simples foi usado para identificar a deleção da glutamato desidrogenase dependente de NADPH e superexpressão da glutamato desidrogenase dependente de NADH, o que resultou em um aumento da produção de etanol, além de uma redução de 40% na geração do subproduto glicerol (NISSEN et al., 2000).

Em outro estudo, um modelo metabólico em escala genômica foi usado para identificar novos genes alvo para manipulação de engenharia genética a fim de melhorar a produção de bioetanol através da inserção de uma gliceraldeído desidrogenase formadora de NADPH, o que resultou em um aumento da produção de etanol com diminuição da formação de glicerol. (BRO et al., 2006).

Recentemente foram produzidos modelos cinéticos que avaliaram fatores que limitam a produção de etanol a partir de xilose através das vias oxi-redutiva e da isomerização na *S. cerevisiae*, verificando que altos níveis de concentração da enzima xiluloquinase são necessários para possibilitar um aumento no consumo de xilose e diminuição da produção de xilitol, o que indica a possibilidade dos modelos

servirem como base para experimentos mais elaborados que possam indicar a direção para aumentar a produção do etanol (PARACHIN et al., 2011).

Em 2007, um grupo de colaboradores pertencentes a diversos grupos de pesquisa, sob a liderança do Centro para Biologia de Sistemas Integrativo de Manchester (MCISB) e a Rede de Biologia de Sistemas de Levedura (YSBN) apresentou uma reconstrução consenso da rede metabólica para *S. cerevisiae*. O trabalho se baseou principalmente nos modelos já existentes iMM904 (MO et al, 2009) e iLL672 (KUEPFER et al, 2005). O modelo consenso resultante do trabalho foi chamado de Yeast 1.0 (HERRGÅRD et al, 2008), e desde então este modelo vem sendo aperfeiçoado pela comunidade científica, chegando à versão 5.0 em 2012 (HEAVNER et al, 2012) e recentemente foram lançadas as versões 6.0 (HEAVNER et al, 2013) e 7.0 (AUNG et al, 2013). A cada versão, o modelo incorpora reações bioquímicas recentemente anotadas para *S. cerevisiae* e busca fazer com que o modelo produza resultados os mais fiéis possíveis à realidade. Esta evolução é apresentada na figura 1.6.

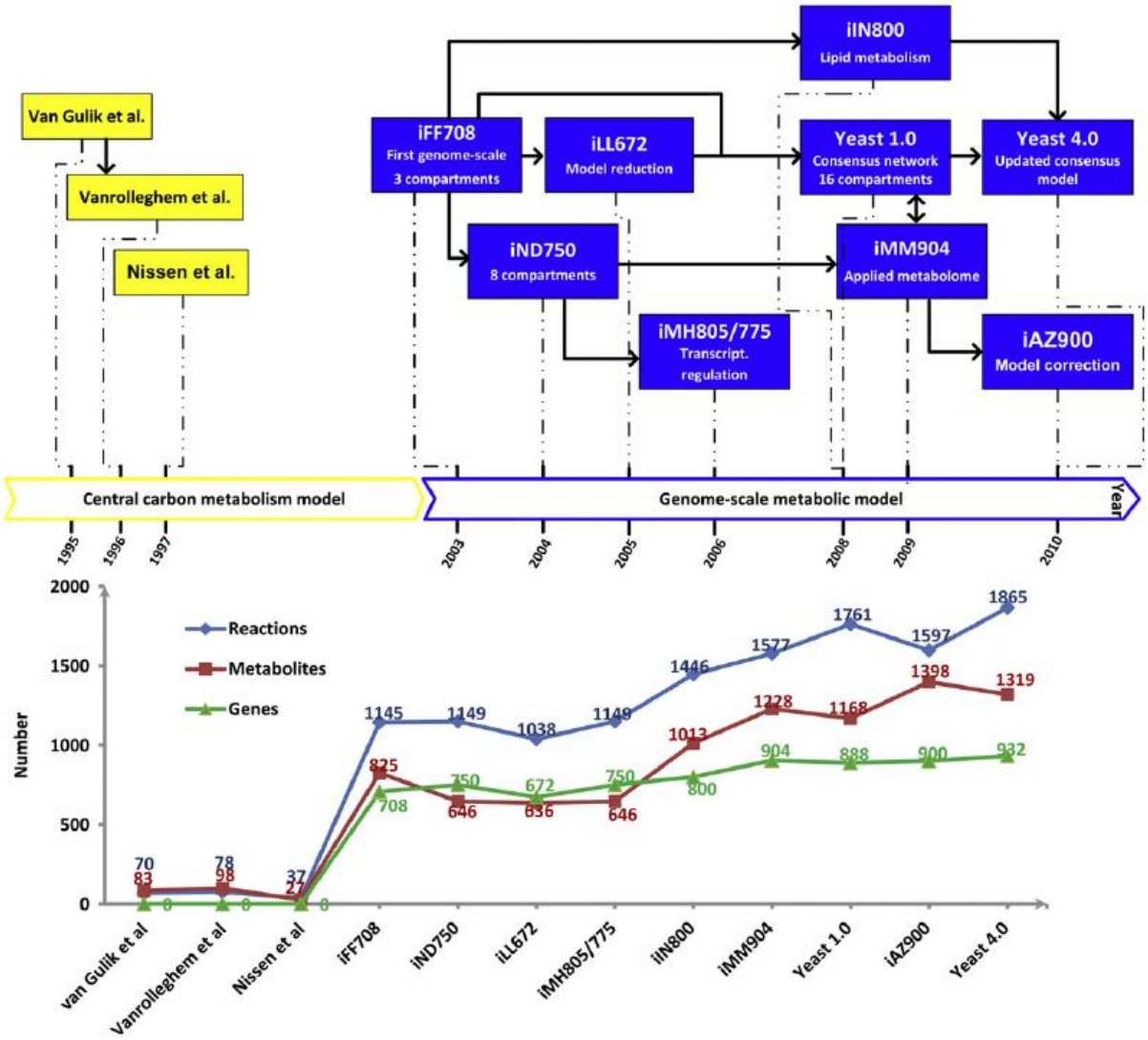


Figura 1.6 (ÖSTERLUND et al., 2012): Evolução dos modelos metabólicos em escala genômica da *S. cerevisiae*, em relação à quantidade de reações, de metabólitos e de genes.

1.6 MODELAGEM ESTRUTURAL DE PROTEÍNAS POR HOMOLOGIA

Outra vertente muito importante da bioinformática é a modelagem estrutural de proteínas por homologia. Com essa metodologia, a sequência de aminoácidos de uma proteína, definida pela sequência de nucleotídeos de um gene, pode ser levada para um contexto tridimensional. As estruturas tridimensionais das moléculas de proteínas são determinadas, sem qualquer contribuição adicional dos ácidos nucleicos, pela sequência de seus aminoácidos. Qualquer enovelamento possível da cadeia principal coloca diferentes resíduos em contato, assim as interações das cadeias laterais e da cadeia principal, consigo mesmas e com o solvente, e as restrições existentes de mobilidade das cadeias laterais, determinam as

estabilidades relativas das diferentes conformações. As proteínas evoluíram de forma que um determinado padrão de enovelamento, chamado de estado nativo, de sua cadeia principal seja termodinamicamente melhor do que outras conformações. (LESK, 2005).

A observação de que cada proteína se enovela espontaneamente em uma conformação nativa tridimensional única implica que a natureza possui um algoritmo para predizer a estrutura de proteínas a partir da sequência de aminoácidos. Algumas tentativas de se entender esse algoritmo são baseadas somente nos princípios gerais da física, que tentam reproduzir as interações interatômicas entre proteínas a fim de definir uma energia mensurável associada a cada conformação. Computacionalmente, o problema da predição de estrutura de proteínas torna-se, então, uma questão de se encontrar o mínimo global desta função de energia conformacional. Até agora estas abordagens não obtiveram êxito, em parte devido à inadequação da função de energia e em parte porque os algoritmos de minimização tendem a ficar presos nos mínimos locais de energia (LESK, 2005).

Uma alternativa a estes métodos por princípios básicos são as abordagens baseadas na pesquisa de dados sobre a estrutura de uma sequência-alvo pela busca de similaridades com estruturas conhecidas. Esses métodos empíricos ou "baseados no conhecimento" estão se tornando bastante eficientes, fornecendo soluções práticas para muitos problemas. O principal exemplo de método empírico é a modelagem por homologia, que prediz a estrutura tridimensional de uma proteína a partir de estruturas conhecidas de uma ou mais proteínas relacionadas. O resultado é um conjunto completo de coordenadas para as cadeias principal e laterais, que dá origem a um modelo estrutural de alta qualidade, comparável a, pelo menos, uma estrutura experimental de baixa resolução. A construção de modelos por homologia é uma técnica útil quando queremos predizer a estrutura de uma proteína-alvo de sequência conhecida e que está relacionada com, pelo menos, outra proteína de sequência e estrutura conhecidas, que pode servir como base para um modelo da proteína-alvo, sendo que a qualidade do modelo depende do grau de similaridade entre as sequências (LESK, 2005).

A modelagem de proteínas por homologia tem sido importante nas áreas de biologia estrutural, bioquímica e biofísica, particularmente em estudos relacionados com os genomas. Aqui, seu potencial é imenso, pois a técnica é capaz de acelerar o

processo de elucidação de estruturas proteicas em curto espaço de tempo e a custos reduzidos. Porém, esta apresenta algumas limitações, como erros na modelagem de cadeias laterais quando o grau de similaridade entre alvo e molde são baixos; distorção na localização de resíduos devido a divergências em regiões das sequências; e erros em regiões da proteína alvo que não existem no molde, ocorrendo normalmente em regiões de alças grandes (SANTOS FILHO; ALENCASTRO, 2003).

1.7 BANCO DE DADOS BIOLÓGICOS

Embora o conhecimento de dados de sequências e estruturas biológicas esteja muito longe de estar completo, ele já apresenta um tamanho respeitável e cresce muito rapidamente. Muitos cientistas trabalham para gerar os dados ou para executar projetos de pesquisa analisando esses resultados. O arquivamento e a distribuição desses dados são realizados por organizações mantenedoras de bancos de dados específicos (LESK, 2005).

Os bancos de dados de sequências geralmente se especializam em um tipo de dados: DNA, RNA ou proteínas. Há grandes coleções de dados de sequências e sites de armazenamento na Europa, no Japão e nos Estados Unidos, e há grupos independentes que espelham todos os dados coletados nos principais bancos de dados públicos, com frequência oferecendo algum *software* que agrega valor aos dados (GIBAS; JAMBECK, 2001).

O Laboratório de Biologia Molecular Europeu (EMBL), o Banco de Dados de DNA do Japão (DDBJ) e os Institutos Nacionais de Saúde dos Estados Unidos (NIH) cooperam, em colaboração com o NCBI, para tornar todos os dados de sequências públicos disponíveis por meio do GenBank, fornecendo a coleção mais completa de dados de sequências de DNA disponível no mundo (GIBAS; JAMBECK, 2001).

Dentre os diversos bancos de dados biológicos disponíveis, três deles são fundamentais para quem trabalha com modelagem estrutural, simulações de via metabólicas e mineração de sequências de proteínas. O PDB (Base de Dados de Proteína) é o principal banco de dados de proteínas, armazenando, além de muitas estruturas tridimensionais de proteínas, estruturas de ácidos nucleicos. A KEGG (Enciclopédia de Genes e Genoma de Kioto) armazena genes, genomas e, principalmente, informações sobre reações das vias metabólicas conhecidas. Por

fim, temos a coleção de sequências proteicas não redundantes mantido pelo NCBI (NR/NCBI) que contém sequências provenientes dos principais bancos de dados de proteínas.

Nesse contexto, o presente trabalho teve como objetivo utilizar ferramentas de biologia de sistema, modelagem estrutural por homologia e banco de dados biológicos de proteínas e vias metabólicas para estudar os problemas de desbalanço de cofatores e não funcionalidade de enzimas diretamente relacionadas com as vias metabólicas exógenas de fermentação de xilose em *S. cerevisiae*.

Na primeira fase do projeto, foram aplicadas ferramentas de análise de balanço de fluxo (FBA) para simular modelos metabólicos em escala genômica de fermentação de xilose por *S. cerevisiae*, a fim de estudar o desbalanço de cofatores da via oxi-redutiva e possíveis alvos de engenharia genética.

Na segunda fase, foram realizados trabalhos de mineração de sequências contidas em bancos de dados de proteínas a fim de encontrar xilose isomerases com potencial de serem funcionalmente expressas em *S. cerevisiae,* além de trabalhos de modelagem por homologia de proteínas funcionais e não funcionais a fim de identificar características estruturais que determinam sua funcionalidade em *S. cerevisiae*.

Capítulo 2

FUNDAMENTAÇÃO TEÓRICA DAS METODOLOGIAS COMPUTACIONAIS UTILIZADAS

2.1 SIMULAÇÕES DE MODELOS METABÓLICOS EM ESCALA GENÔMICA

2.1.1 ANÁLISE DE BALANÇO DE FLUXO (FBA) E MINIMIZAÇÃO DE AJUSTE METABÓLICO (MOMA)

FBA é uma abordagem matemática amplamente utilizada para analisar os fluxos de metabólitos em uma rede metabólica presumidamente em estado estacionário, ou seja, uma rede onde os fluxos das reações internas são constantes e por consequência o fluxo de metabólitos consumidos é igual ao fluxo de metabólitos produzidos. Biologicamente, é considerado que um organismo opera próximo do estado estacionário quando este apresenta taxa de crescimento constante e taxa de variação dos metabólitos internos desprezíveis.

Para realizar as análises, a rede metabólica é normalmente representada na forma de um modelo metabólico em escala genômica, que consiste em uma matriz numérica contendo os coeficientes estequiométricos de todas as reações conhecidas do organismo e os genes que codificam cada enzima. Ou seja, na matriz estequiométrica, cada linha representa uma reação da rede metabólica, que pode ser biológica, química ou de transporte, sendo que o modelo não fará diferença entre elas. Já cada coluna da matriz representa um metabólito participante da rede, podendo ser consumido, produzido ou simplesmente um intermediário. O sistema pode ser representado pela equação 2.1:

$$S \times v = 0 \qquad (2.1)$$

, onde S é a matriz estequiométrica e v é o vetor de fluxos, que serão determinados através do cálculo do FBA.

Pala ilustrar os primeiros passos da montagem matemática do FBA, será utilizado como exemplo o sistema da figura 2.1 que representa uma célula formada por apenas 7 genes (3 transportadores e 4 enzimas, representando 5 reações irreversíveis e 1 reversível) e 3 metabólitos (A, B e C).

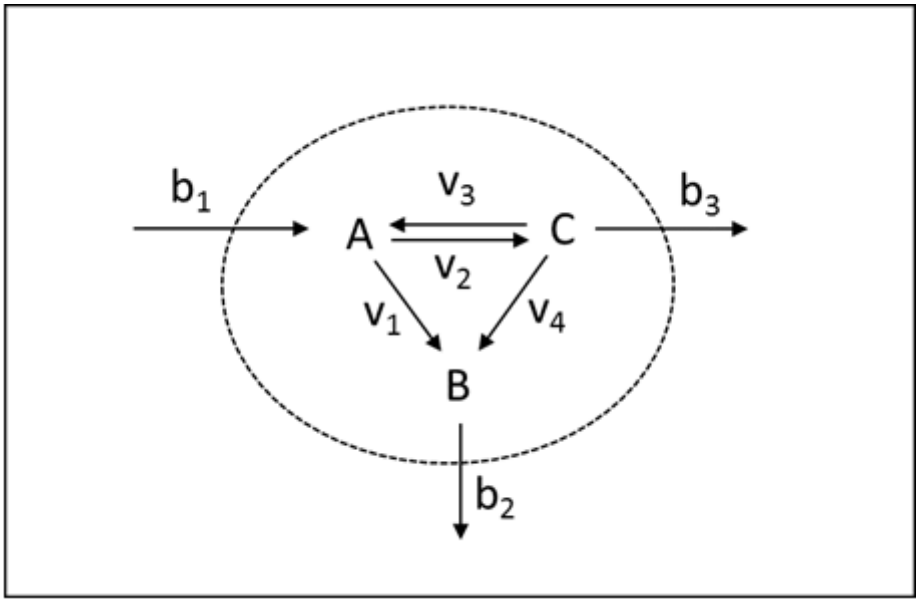


Figura 2.1: Sistema representando um organismo hipotético com 3 metabólitos (A, B e C) e 6 genes, sendo 3 transportadores: b_1 – representa o fluxo da reação responsável por absorver o metabólito A, b_2 – representa o fluxo da reação responsável por secretar o metabólito B e b_3 – representa o fluxo da reação responsável por secretar o metabólito C; e 3 enzimas: v_1 – representa o fluxo da reação que converte A em B, v_2/v_3 – representam o fluxo da reação que converte A em C, sendo que esta é uma reação reversível e v_4 – representa o fluxo da reação que converte C em B.

Ao aplicar balanço de massa de estado estacionário para os 3 metabólitos (A, B e C) obtemos o conjunto de equações 2.2:

$$\begin{aligned} &\text{A:} && -v_1 - v_2 + v_3 + b_1 = 0 \\ &\text{B:} && v_1 + v_4 - b_2 = 0 \\ &\text{C:} && v_2 - v_3 - v_4 - b_3 = 0 \end{aligned} \tag{2.2}$$

Sendo que o sinal negativo indica que o metabólito está sendo consumido e o sinal positivo que ele está sendo gerado. Colocando-se as equações na forma matricial obtêm-se as matrizes descritas em 2.3 e 2.4:

$$S = \begin{bmatrix} -1 & -1 & 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & -1 & 0 \\ 0 & 1 & -1 & -1 & 0 & 0 & -1 \end{bmatrix} \quad (2.3)$$

$$v = [v_1 \quad v_2 \quad v_3 \quad v_4 \quad b_1 \quad b_2 \quad b_3]^T \quad (2.4)$$

Desse modo, temos um sistema linear com 3 equações e 7 incógnitas, caracterizando, portanto, um sistema indeterminado. Os coeficientes de S impõem restrições aos fluxos dos metabólitos da rede, garantindo que a quantidade total de qualquer composto produzido será igual à sua quantidade total sendo consumida. A cada reação também podem ser adicionados limites superiores e inferiores que definem os fluxos máximos e mínimos permitidos para a reação. Esses valores são baseados em conhecimentos biológicos do organismo a ser estudado e são fundamentais para manter os resultados finais dentro do esperado biologicamente. O conjunto de todas as restrições define o espaço de distribuições de fluxo permitidas no sistema, ou seja, as taxas com que cada reação pode produzir ou consumir cada metabólito (ORTH et al, 2010).

Para o exemplo acima, podem ser definidas as seguintes restrições descritas nas relações matemáticas em 2.5:

$$\begin{aligned} &v_1 \geq 0, v_2 \geq 0, v_3 \geq 0, v_4 \geq 0, \\ &b_1 \geq 0, b_1 \leq 10, b_2 \geq 0, b_3 \geq 0 \end{aligned} \quad (2.5)$$

Note que as restrições obrigam um sentido único de cada fluxo e também limitam o fluxo de consumo do metabólito A em 10 unidades, sem essa restrição a FBA pode vir a predizer um crescimento infinito. A limitação de fluxos para v_1 e v_4 para maiores ou iguais a zero vem da observação biológica de que essas reações hipotéticas são irreversíveis, enquanto, como a reversibilidade da reação de v2 e v3 já está representada com a presença de dois fluxos inversos, estes também devem ser maiores ou iguais a 0, note que esta reação poderia ser representada por apenas um fluxo não limitado. A escolha de como representar a reação, neste caso, não influencia nos resultados das simulações futuras.

Apesar da inclusão de restrições, o sistema ainda admite infinitas soluções e, portanto, para determinar qual das soluções é a que reflete o funcionamento biológico, a abordagem FBA utiliza o conceito de programação linear para encontrar

a solução desejada. Dessa forma, para encontrar a solução “ótima” de um sistema indeterminado é necessário definir uma “função objetivo” que deverá ser maximizada ou minimizada. Essa função pode ser uma das equações já existentes no sistema (otimizando uma das variáveis em relação as outras), ou uma nova função que utiliza as variáveis já definidas. Assim, para realizar a FBA é preciso definir um objetivo biológico que é relevante ao problema estudado e no caso da predição de crescimento, o objetivo é maximizar a produção de biomassa, definida através da taxa de produção dos compostos metabólicos que vão contribuir para a formação da biomassa. Matematicamente isso é representado pela função objetivo, que indica quanto cada reação do modelo contribui com o fenótipo estudado (ORTH et al, 2010).

Supondo, de acordo com o exemplo descrito acima, que a biomassa do organismo modelado é composta por 60% do metabólito B e 40 % do metabólito C, dessa forma a função objetivo pode ser representada pela equação 2.6:

$$Z = 0{,}6\ b2 + 0{,}4\ b3 \tag{2.6}$$

Ou na forma matricial descrita em 2.7:

$$\begin{aligned} & c^t \times v = 0 \\ & c^T = [0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0{,}6 \quad 0{,}4] \\ & v = [v_1 \quad v_2 \quad v_3 \quad v_4 \quad b_1 \quad b_2 \quad b_3]^T \end{aligned} \tag{2.7}$$

Em conjunto, as representações matemáticas das reações (2.4 e 2.5) e a função objetivo (2.7) definem um sistema linear de equações que é normalmente possível e indeterminado, permitindo infinitas soluções. Busca-se então a solução que maximiza ou minimiza (dependendo do que é estudado) a função objetivo, o que representaria a realidade do organismo estudado. Esta solução é encontrada utilizando-se algoritmos de programação linear, que conseguem achar a solução ótima rapidamente mesmo para grandes sistemas (ORTH et al, 2010). O passo-a-passo resumido desta abordagem está resumido na figura 2.2.

A análise de FBA pode ser usada também para simular o comportamento de

um organismo geneticamente modificado, simplesmente alterando a(s) reação(ões) que representa(m) a mutação desejada, seja eliminando a reação do modelo (equivalente a deleção de um gene), adicionando uma nova reação (inserção de um novo gene) ou alterando seus limites (mudança na taxa de expressão da proteína através da mudança do promotor).

Entretanto essa abordagem também tem suas limitações, uma vez que não utilizam de parâmetros cinéticos por serem muito difíceis de serem obtidos experimentalmente, a análise de FBA não prediz a concentração dos metabólitos, sendo também adequada apenas para predizer os fluxos no estado estacionário. A FBA também não considera efeitos regulatórios como ativação de enzimas através de eventos de fosforilação causados por quinases ou regulação da expressão de genes.

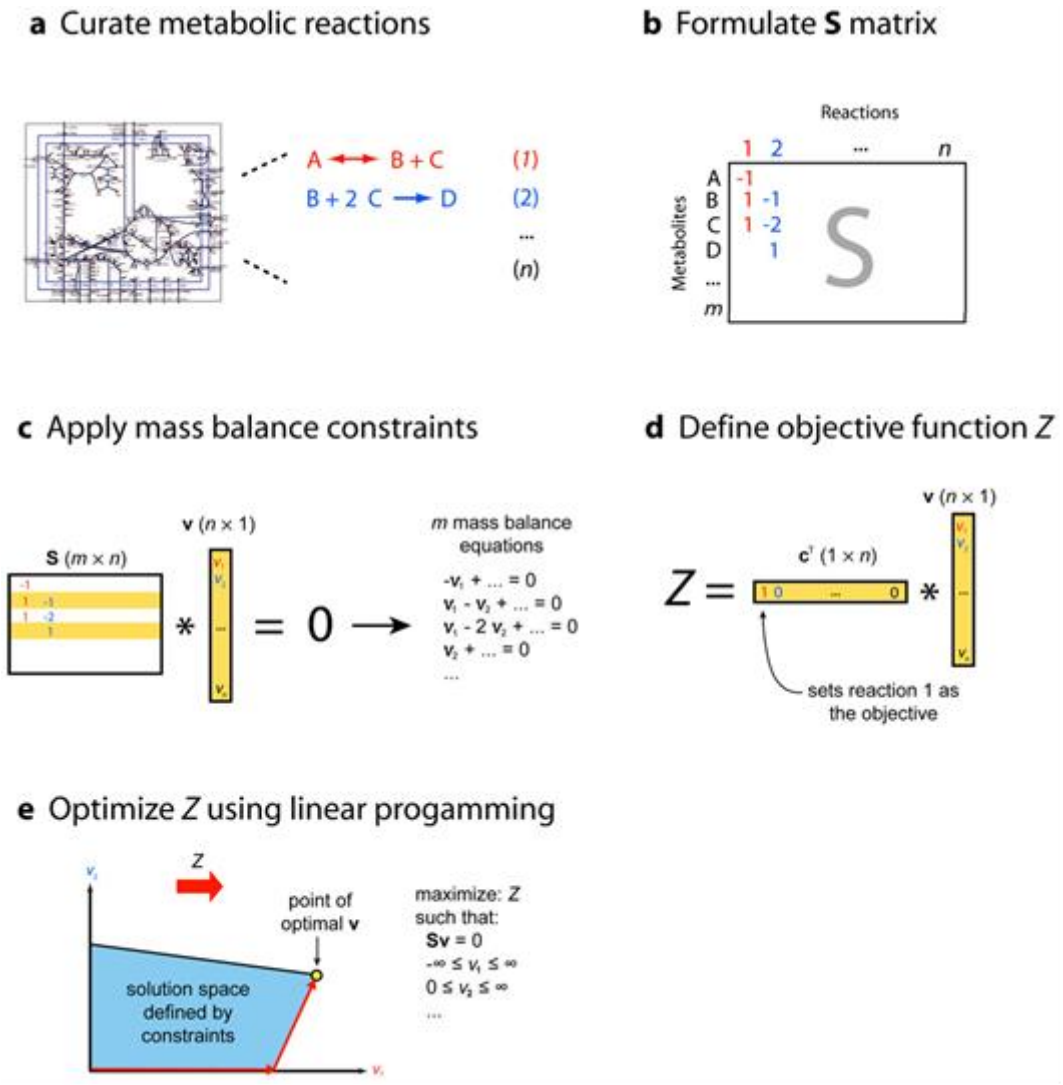


Figura 2.2 (ORTH JD et al, 2010): Passo-a-Passo da FBA. (a) É obtida uma lista mais completa e curada possível contendo as reações metabólicas de um organismo. (b) A partir da estequiometria dessas reações, é elaborada uma matriz matemática onde cada linha representa um metabólito e cada coluna uma reação. (c) Como o balanço de massa deve ser mantido, o produto da matriz pelos fluxos em cada reação deve ser igual a 0. Obtemos assim um conjunto de equações que podem ser suplementadas por outros limites observados no organismo (reações irreversíveis, limite de substrato disponível, etc.), chegando assim a um sistema linear que normalmente tem infinitas soluções. (d) É definida uma função objetivo (na figura vemos que a reação 1 é escolhida como objetivo). (e) A FBA utiliza programação linear para encontrar a solução que maximiza ou minimiza a função objetivo.

Além disso, a FBA pode não refletir a realidade em casos de organismos modificados, já que ela sempre apresentará a solução que otimiza o objetivo dado, mesmo que esta seja muito diferente da solução identificada no organismo selvagem (produção de biomassa, por exemplo). Para contornar essa questão, foi desenvolvida a minimização de ajuste metabólico (MOMA), uma abordagem

baseada em FBA que busca no espaço de soluções factíveis do organismo modificado aquela que minimiza a distância entre ela e a solução dada pela FBA para o organismo selvagem, utilizando-se de programação quadrática. Sua lógica está representada na figura 2.3.

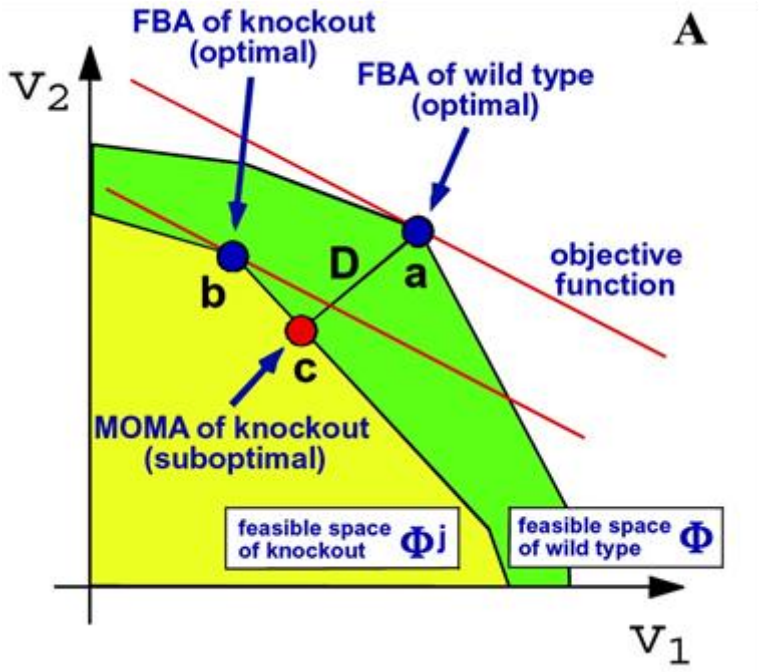


Figura 2.3 (SEGRÈ et al, 2002): Lógica usada pela MOMA. A área verde representa o espaço de solução possível do modelo selvagem, enquanto a amarela representa o espaço possível do modelo mutante. A solução que maximiza a função objetivo é indicada pelo ponto “a” para o modelo selvagem e pelo ponto “b” para o modelo mutante, esses dois pontos são os encontrados pela abordagem FBA. O ponto “c” representa o ponto dentro do espaço de possíveis soluções do mutante que está mais próximo do ponto “a”, sendo encontrado pela MOMA.

2.1.2 SOFTWARES UTILIZADOS PARA AS ANÁLISES DE FBA

As simulações foram realizadas pelo software MATLAB em sua versão R2011a (MATHWORKS, 2011) utilizando o pacote de biologia de sistemas COBRA toolbox 2.0.5 (SCHELLENBERGER et al., 2011) e o solver Gurobi 5.6.0 (GUROBI OPTIMIZATION, 2004), necessário principalmente para resolver a programação quadrática utilizada pela MOMA.

A abordagem COBRA (Reconstrução e Análise Baseadas em Restrições) vem sendo amplamente utilizada na comunidade científica na última década, consistindo em uma série de métodos para simular, analisar e predizer fenótipos

metabólicos de modelos em escala genômica, dentre esses métodos estão presentes as abordagens FBA e MOMA. A seguir é apresentada a tabela 2.1, com algumas funções importantes para o trabalho fornecidas pela ferramenta COBRA Toolbox, um pacote para MATLAB que implementa os métodos COBRA, e também a figura 2.4, com uma visão geral de todos os pacotes e funções da ferramenta.

Tabela 2.1: Métodos do COBRA toolbox importantes para a pesquisa

Método	**Descrição**
readCbModel	Lê o modelo no formato *SBML* (como é distribuído) e o armazena em uma estrutura do MATLAB
changeRxnBounds	Altera os limites inferior e/ou superior de determinada reação
addReaction	Adiciona uma nova reação ao modelo
removeRxns	Remove uma reação do modelo (pode ser substituída por *changeRxnBounds* fixando ambos os limites em 0)
printRxnFormula	Imprime a fórmula de uma determinada reação
optmizeCbModel	Utiliza FBA para simular o modelo e armazena o resultado em uma estrutura do MATLAB
MOMA	Utiliza MOMA para simular o modelo e armazena o resultado em uma estrutura do MATLAB (necessita de um solucionador de programação quadrática e de um modelo que represente a “linhagem selvagem”)

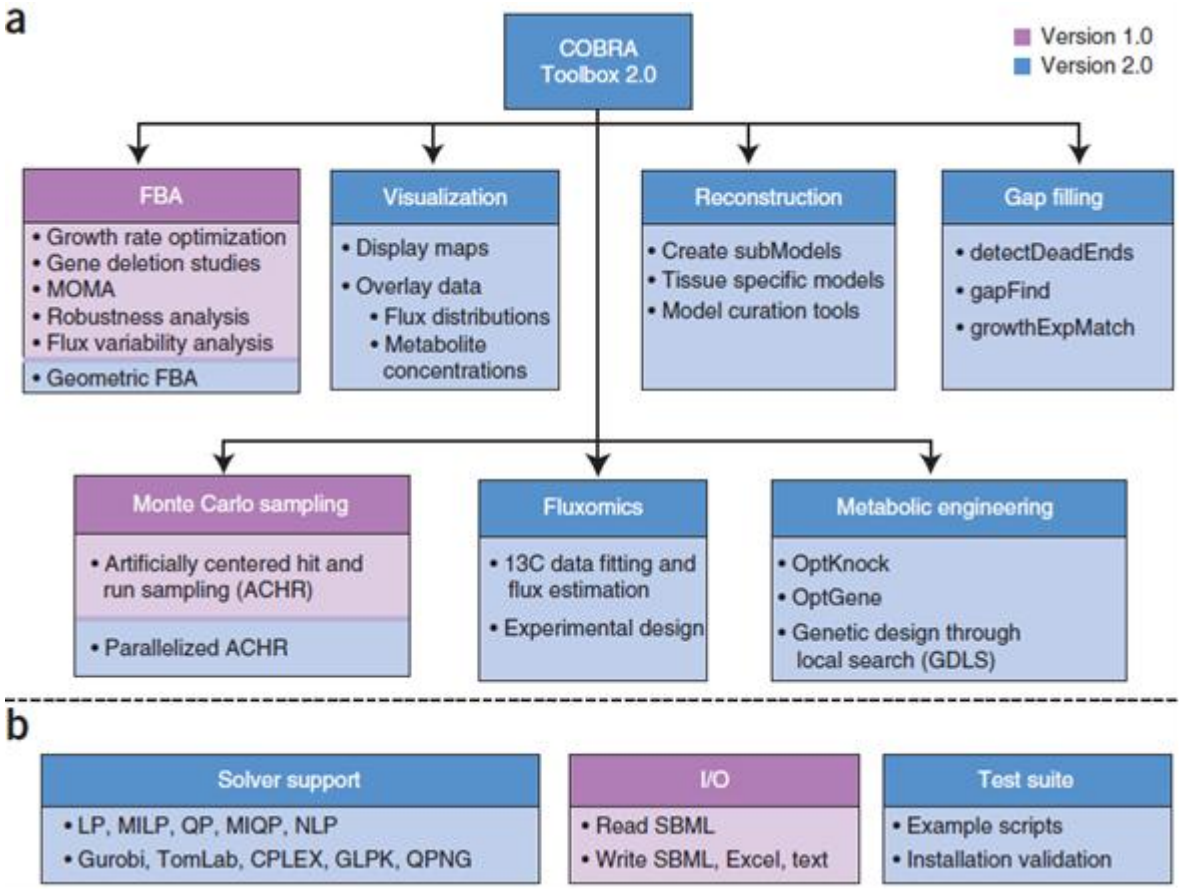


Figura 2.4 (SCHELLENBERGER et al., 2011): Visão geral dos pacotes do COBRA Toolbox

A versão do modelo metabólico de levedura *Yeast 5.0* foi escolhida como a base para as simulações de fluxo. Ele é formado por 1418 metabólitos e 2110 reações, sendo 170 reações de troca de metabólitos com o meio (produção ou consumo). O modelo como é distribuído define um meio aeróbico com consumo de glicose limitado, e permite trocas de oxigênio, amônio, prótons, ferro (2+), fosfato, potássio, sódio, sulfato e água de modo ilimitado com o meio. Para simular crescimento anaeróbico, a reação de troca de oxigênio com o meio deve ser limitada de modo a não permitir o consumo do mesmo, além disso, a simulação de crescimento anaeróbico deve permitir trocas ilimitadas de ergosterol, lanosterol, zimosterol e fosfatidato com o meio, e a definição de biomassa deve ser alterada removendo-se 14-demetillanosterol e ergosta-5,7,22,24(28)-tetraen-3beta-ol da definição de “lipídio”. Esta necessidade reflete a observação experimental que leveduras necessitam de esteróis e ácidos graxos quando cultivados sob condições anaeróbicas rígidas. Entretanto, de uma perspectiva de modelagem essas necessidades surgem devido à definição de biomassa, para a qual a bioquímica

ainda não está bem estabelecida, e à reconstrução do metabolismo de esterol, o qual está incompleto no modelo *Yeast 5.0* (HEAVNER et al, 2012).

2.2 MODELAGEM POR HOMOLOGIA E ANÁLISE ESTRUTURAL DE PROTEÍNAS

2.2.1 MODELAGEM DE PROTEÍNAS POR HOMOLOGIA

No método de modelagem por homologia, também chamada de modelagem comparativa, a proteína de interesse (proteína alvo) terá sua estrutura tridimensional predita usando como referência a estrutura de outra proteína similar (também chamada de molde), na maioria das vezes evolutivamente relacionada com o alvo. Essa proteína deve possuir estrutura 3D resolvida experimentalmente, e as coordenadas cartesianas de seus átomos devem estar depositadas em banco de dados de estruturas, como o PDB. A modelagem por homologia é o método empregado mais frequentemente, e sua capacidade de predição está intrinsecamente relacionado com o grau de similaridade entre as estruturas alvo e molde. Geralmente, consideram-se como limites mínimos de aplicabilidade do método valores de 25 a 30% de identidade, obtidos através do alinhamento global entre a estrutura primária da proteína alvo e de uma ou mais proteínas molde. (CAPRILES et al., 2014). A seguir serão apresentadas as etapas comumente seguidas por um experimento de modelagem por homologia (As explicações das etapas foram baseadas em: SAXENA et al, 2013; KRIEGER et al, 2003; SANTOS FILHO; ALENCASTRO, 2002; LESK, 2005; Documentação do programa YASARA).

Seleção das proteínas-molde

O ponto inicial em uma modelagem por homologia é identificar as proteínas de estrutura conhecidas relacionadas à proteína-alvo e selecionar aquelas que serão usadas como molde. Um dos métodos mais comuns é a geração de um perfil da proteína alvo através de alinhamento múltiplo e utilização deste para buscas de proteínas relacionadas depositadas em bases de dados, como o PDB. A partir da lista de proteínas obtidas pela busca, é necessário selecionar aquelas apropriadas

para servir de molde da modelagem, sendo que cada molde dará origem a um modelo. Normalmente, quanto maior a semelhança entre alvo e molde, melhor o modelo produzido, mas, além disso, outros fatores também são levados em conta para realizar a seleção, como maior proximidade filogenética, existências das proteínas em ambientes semelhantes e qualidade do modelo da proteína-molde.

Alinhamento entre alvo e molde

O próximo passo consiste no alinhamento entre alvo e cada molde. Apesar da fase anterior gerar alinhamentos, estes normalmente não são ideais para a modelagem, pois são ajustados para identificação de proximidade filogenética e não para produzir alinhamentos otimizados. Logo, é utilizado um método otimizado para alinhamento global entre a sequência-alvo e a estrutura do modelo, sendo comum o uso de algoritmos de programação dinâmica que utilizam matrizes de substituição como PAM e BLOSUM. Para proteínas com relação próxima com identidade maior que 40%, o alinhamento é praticamente sempre correto, sendo que regiões com baixa similaridade se tornam comum em alinhamentos de identidade menor que este valor. Normalmente, pode ser observado que inserções ou deleções ocorrem nas regiões de alças, entre elementos de estrutura secundária regular, como hélices e fitas.

Modelagem da cadeia principal

A partir do alinhamento, o esqueleto da proteína é criado de maneira trivial: copiando-se as coordenadas dos resíduos do molde que se alinharam com a sequência-alvo. No caso de os resíduos alinhados serem diferentes, apenas as coordenadas da cadeia principal são copiadas, caso contrário a cadeia lateral também pode ser incluída. Se existirem moldes em estados oligoméricos, os modelos podem ser construídos no mesmo estado, de modo que interações entre cadeias laterais através da interface podem ser consideradas.

Modelagem das alças

A modelagem das alças é uma etapa essencial no processo de criação do modelo, pois elas, sendo as regiões mais expostas da proteína, sofrem frequentes

mutações, o que determina a especificidade das proteínas. Isso causa, como já dito anteriormente, as inserções e deleções observadas no alinhamento.

A modelagem das alças pode ser realizada de duas formas: método *ab initio*, que utiliza minimização de energia para predizer o enovelamento da alça; e o método comparativo, mais utilizado e normalmente com melhores resultados, que utiliza bases de dados, como o PDB, para buscar alças conhecidas que se encaixem na região a ser predita.

Modelagem das cadeias laterais

Nas regiões onde os resíduos alinhados diferem é necessário realizar a predição das cadeias laterais. Para isso, são usadas bibliotecas de conformações comuns, ou rotâmeros, de cadeias laterais. Essas bibliotecas indicam as possíveis conformações que a cadeia lateral pode ter em função dos ângulos de conformação da cadeia principal.

Otimização do modelo

Nesta fase, o objetivo é integrar as três modelagens anteriores. As alças são otimizadas testando-se uma grande quantidade de conformações diferentes e reotimizando as cadeias laterais para cada uma delas, enquanto os rotâmeros da cadeia lateral são otimizados considerando-se interações eletrostáticas, interações baseadas em conhecimento e efeitos de solvatação. Por fim, é realizada uma minimização de energia de alta resolução utilizando campos de força baseados em conhecimento.

Validação e construção de modelo híbrido

Após a modelagem a partir de cada proteína-molde, indicadores de qualidade para os modelos resultantes são determinados baseados na qualidade do molde, qualidade estereoquímica, interação entre a estrutura e o meio, mecânica molecular, cálculos de energia livre, entre outros.

Após a validação, um modelo híbrido pode ser construído, de modo que regiões de baixa qualidade provenientes dos modelos com melhores indicadores são iterativamente substituídas por fragmentos correspondentes de outros modelos.

2.2.2 SOFTWARE UTILIZADO PARA MODELAGEM ESTRUTURAL

YASARA (http://www.yasara.org/) é um programa utilizado para modelagem, simulação e visualização gráfica de moléculas. Pode ser usado tanto em modo gráfico quanto modo texto, possuindo uma linguagem de macros chamada *Yanaconda*. O programa também disponibiliza um pacote com todas as macros como funções na linguagem Python, facilitando a construção de *scripts*. A seguir são apresentadas algumas definições utilizadas pelo YASARA que são importantes para entender suas macros (Definições retiradas da documentação do YASARA):

- **Átomo**: Cada átomo possui um identificador numérico único e várias propriedades, como elemento químico, nome, nome do resíduo, número do resíduo, nome da molécula, etc. Cada átomo se refere a uma linha de um arquivo PDB.
- **Resíduo**: Um trecho contínuo de átomos que possuam o mesmo nome e número de resíduo e o mesmo nome de molécula.
- **Molécula**: Trecho contínuo de resíduos que compartilham o mesmo nome de molécula.
- **Objeto**: Coleção de moléculas e outros itens, como rótulos e setas. Podem ser ativos ou inativos, quando não são exibidos e nem participam de simulações.
- **Sopa**: Conjunto de todos átomos pertencentes a objetos ativos
- **Cena**: Conjunto de todos objetos, ativos e inativos

A seguir serão apresentadas as principais funções utilizadas na pesquisa (Material baseado na documentação do YASARA).

Modelagem por Homologia

A função *ExperimentHomologyModeling* realiza a modelagem por homologia de uma proteína – seguindo os passos explicados na seção anterior – a partir de um arquivo contendo a sequência a ser modelada e modelos tridimensionais já existentes, providos pelo usuário ou existentes no PDB. O método produz como resultado um modelo híbrido baseado nos melhores modelos construídos.

Minimização e Cálculo de Energia

Após a modelagem da proteína, é realizado um experimento de minimização com a função *ExperimentMinimization*, a fim de remover superposições e corrigir eventuais posicionamentos incorretos. Primeiramente, o experimento corrige estresses conformacionais com uma curta minimização utilizando o método do gradiente; depois é utilizado *Simulated annealing* até ser atingido um critério de convergência da energia (calculada a cada 200 passos, sendo que cada passo representa 2fs da simulação). A convergência é detectada quando a energia tiver uma alteração menor que um valor de corte (valor padrão: 0.05 kJ/mol) por átomo em relação ao passo anterior.

Para calcular a energia necessária para se separar dois objetos, ou seja, a energia de ligação, é utilizado o método *BindEnergy*. Este método recebe como parâmetro um objeto e calcula a energia de ligação entre ele e o restante da sopa, dado um determinado campo de força. Para tanto, ele calcula a energia do sistema considerando os elementos a uma distância infinita um do outro (estado desvinculado) e subtrai a energia do sistema considerando os objetos na distância normal em que estão (estado vinculado). Quanto mais positiva essa energia, mais favorável será a interação entre os objetos, sendo mais difícil separá-los.

A função *ForceField* determina qual campo de força será utilizado em simulações em experimentos de minimização e cálculos de energia de ligação. No caso desta pesquisa, foi utilizado o campo YAMBER3 (KRIEGER et al., 2004), o mais atual da família YAMBER, que são campos de força otimizados para manter a qualidade estrutural e melhorar modelos de homologia. Uma das principais vantagens destes campos de força é a implementação de atrações de Van der Waals mais fortes, fazendo com que espaços vazios entre as cadeias laterais, que tendem a ser formados durante uma modelagem por homologia, sejam removidos, melhorando o modelo.

Alinhamento estrutural de proteínas

O método *AlignMol* dá a opção de se escolher um algoritmo entre um conjunto de opções para realizar o alinhamento estrutural de proteínas. Esta função realiza o alinhamento estrutural de duas proteínas e retorna:

- Os valores RMSD dos resíduos alinhados
- A identidade do alinhamento
- O número de resíduos alinhados
- Os pares de resíduos alinhados

Para o presente estudo, os alinhamentos foram realizados utilizando-se o método MUSTANG (Algoritmo de Alinhamento Estrutural Múltiplo), um algoritmo robusto que se utiliza de informação espacial dos carbonos alfa para construir o alinhamento de múltiplas estruturas de proteínas (KONAGURTHU et al., 2006).

Capítulo 3

SIMULAÇÕES DE MODELOS METABÓLICOS EM ESCALA GENÔMICA

A fim de analisar o fluxo de metabólitos durante a fermentação de xilose na *S. cerevisiae*, e encontrar possíveis alterações que levem a uma maior eficácia na produção de etanol, a abordagem FBA foi utilizada para realizar simulações sobre um modelo metabólico em escala genômica de *S. cerevisiae* com a inclusão das vias oxi-redutiva e de isomerização.

Este processo envolveu o estudo de métodos e ferramentas computacionais para se realizar simulações em um modelo metabólico, a escolha do modelo de *S. cerevisiae* mais adequado para o projeto, testes de robustez baseados em dados experimentais da literatura que exploraram eventos de superexpressão e deleção de genes, elaboração de métodos computacionais para apoiar as simulações e, por fim, a análise dos resultados das simulações.

3.1 ESTUDOS PRELIMINARES

Primeiramente, foram realizados estudos preliminares a fim de se familiarizar com os algoritmos da simulação e também para avaliar a robustez do modelo utilizando resultados experimentais descritos na literatura. Nesta fase, simulou-se o crescimento aeróbico e anaeróbico de *S. cerevisiae* a partir do consumo de glicose e avaliou-se a capacidade de predição do modelo com a simulação de resultados experimentais de um estudo sobre aumento da produção de glicerol através de deleção e superexpressão de genes (CORDIER et al., 2007). Também durante esta fase foram desenvolvidos dois *scripts* em MATLAB, a fim de flexibilizar e automatizar as simulações.

3.1.1 *SCRIPTS* DESENVOLVIDOS PARA OTIMIZAÇÃO DAS ANÁLISES

O próprio artigo do *Yeast 5.0* (HEAVNER et al, 2012) já provê alguns *scripts* que facilitam a simulação do modelo utilizando o pacote do COBRA toolbox, a citar, o *script fluxDistribution.m* provê os comandos necessários para realizar a simulação dos

crescimentos aeróbico e anaeróbico da *S. cerevisiae* e exibir, além do fluxo de biomassa produzida, os fluxos de algumas reações que são especificadas em uma lista no código fonte do *script*, sendo que esta lista pode ser alterada para incluir qualquer outra reação de interesse.

Logo no início do trabalho, surgiu a necessidade de desenvolver dois novos *scripts*, sendo um para facilitar a criação de novos modelos derivados do modelo padrão *Yeast 5.0* (*addModels.m*) e outro para realizar diversas simulações diferentes em sequência, comparando os fluxos entre elas e apresentando apenas as diferenças significativas de forma a facilitar a interpretação dos novos fenótipos gerados pelos modelos (*diffModels.m*). A seguir, o funcionamento desses dois *scripts* será apresentado.

addModels.m

O *script addModels.m* recebe dois parâmetros: uma estrutura do MATLAB representando o modelo padrão (modelo base que no nosso caso representa o *Yeast 5.0*) já previamente lido pelo MATLAB (através da função *readCbModel*) e o nome de um arquivo no formato planilha do excel *".xls"* (Figura 3.1) podendo conter várias planilhas, no qual, em cada planilha, são descritas as alterações a serem feitas no modelo padrão para a criação de um novo modelo, sendo que cada linha representa a alteração de uma reação do modelo (mudança de limites, adição ou exclusão de reação). O *script* consiste em ler cada linha de cada planilha; modificar a reação indicada, utilizando a função do COBRA *addReaction* para adicionar uma nova reação e a função *changeRxnBounds* para alterar os limites ou excluir (alterar os limites para 0) de uma reação; e salvar os novos modelos em uma estrutura do MATLAB que manterá todos os modelos a serem trabalhados.

A criação desse *script* foi importante para facilitar a análise em larga escala de vários modelos além de permitir que outros usuários sem muito conhecimento de programação em MATLAB possam realizar suas próprias simulações.

	A	B	C	D	E
1	**Operacao**	**Reacao**	**Lower Bound**	**Upper Bound**	**Formula**
2	chg	oxygen exchange	0		
3	chg	D-glucose exchange	0		
4	chg	lipid pseudoreaction [no 14-demethyllanosterol, no ergosta-5,7,22,24(28)-tetraen-3beta-ol]	-Inf	Inf	
5	chg	ergosterol exchange	-Inf	Inf	
6	chg	lanosterol exchange	-Inf	Inf	
7	chg	zymosterol exchange	-Inf	Inf	
8	chg	phosphatidate exchange	-Inf	Inf	
9	chg	lipid pseudoreaction		0	
10	chg	D-xylose exchange	-1000		
11	add	r_XI			s_0578 <=> s_0580
12	NOTA				D-xylose <=> D-xylulose
13	NOTA	Nome: X_XI			

Figura 3.1: Planilha “xls” com alterações para permitir o crescimento anaeróbio com consumo de xilose e inclusão da enzima xilose isomerase. Neste exemplo, podemos notar que as linhas, de cima para baixo, impedem o consumo de oxigênio e glicose, liberam as trocas dos compostos necessários para haver fermentação, “trocam” a definição de pseudorreação de lipídio (linhas 4 e 9), libera o consumo de até 1000 mmol/gDW/h de xilose e insere a reação da enzima xilose isomerase, que transforma xilose em xilulose (Observação: a primeira linha e as iniciadas por “NOTA”, são ignoradas pelo *script*).

diffModels.m

O *script diffModels.m* recebe os seguintes parâmetros: os modelos a serem simulados (previamente construídos e lidos pelo MATLAB), as abordagens a serem usadas para simular cada modelo (FBA e/ou MOMA) e dois valores de corte utilizados para imprimir uma reação caso exista uma diferença significativa entre seus fluxos em simulações de modelos diferentes: o primeiro valor define uma diferença mínima entre os fluxos enquanto o segundo estabelece uma razão mínima. Além disso, também podem ser passadas como parâmetro reações que devem ser impressas mesmo se não atingirem os níveis desejados de diferença. Também, no caso de se desejar utilizar a abordagem MOMA para a simulação de algum modelo, é necessário indicar qual modelo será usado como o “modelo selvagem” do algoritmo.

Após executar as funções *optmizeCbModel* e/ou *MOMA* para cada modelo, é feita uma lista de reações que obedecem aos critérios passados como parâmetros, em seguida essa lista é percorrida, imprimindo de forma tabular as informações relevantes de cada reação, como nome da reação, código desta no modelo, equação estequiométrica da reação e fluxo sobre a mesma para cada modelo testado.

O desenvolvimento desse *script* foi fundamental para auxiliar na correta interpretação dos novos fenótipos obtidos pelas simulações, pois pequenas mudanças em restrições ou reações podem levar a grandes mudanças de fluxos que precisam ser interpretadas do ponto de vista biológico.

A seguir serão apresentados os resultados das simulações utilizando dados experimentais retirados da literatura com o objetivo de avaliar as metodologias desenvolvidas.

3.1.2 SIMULAÇÃO DE CRESCIMENTOS AERÓBICO E ANAERÓBICO UTILIZANDO GLICOSE COMO SUBSTRATO

Para o primeiro teste, o *script fluxDistribution.m* foi alterado a fim exibir algumas reações para fins de comparação entre um crescimento aeróbico (respiração) e um crescimento anaeróbico (fermentação) tendo como substrato glicose, que teve um fluxo limitado a 1000 mmol/gDW/h para ambas simulações, sendo que para a respiração o consumo de oxigênio permaneceu ilimitado e para a fermentação este foi zerado. O gráfico comparando os fluxos de consumo e produção de alguns metabólitos entre um modelo aeróbio e um anaeróbio é apresentado na figura 3.2. O modelo *Yeast 5.0* define que valores negativos e positivos significam, respectivamente, consumo e produção de determinado metabólito.

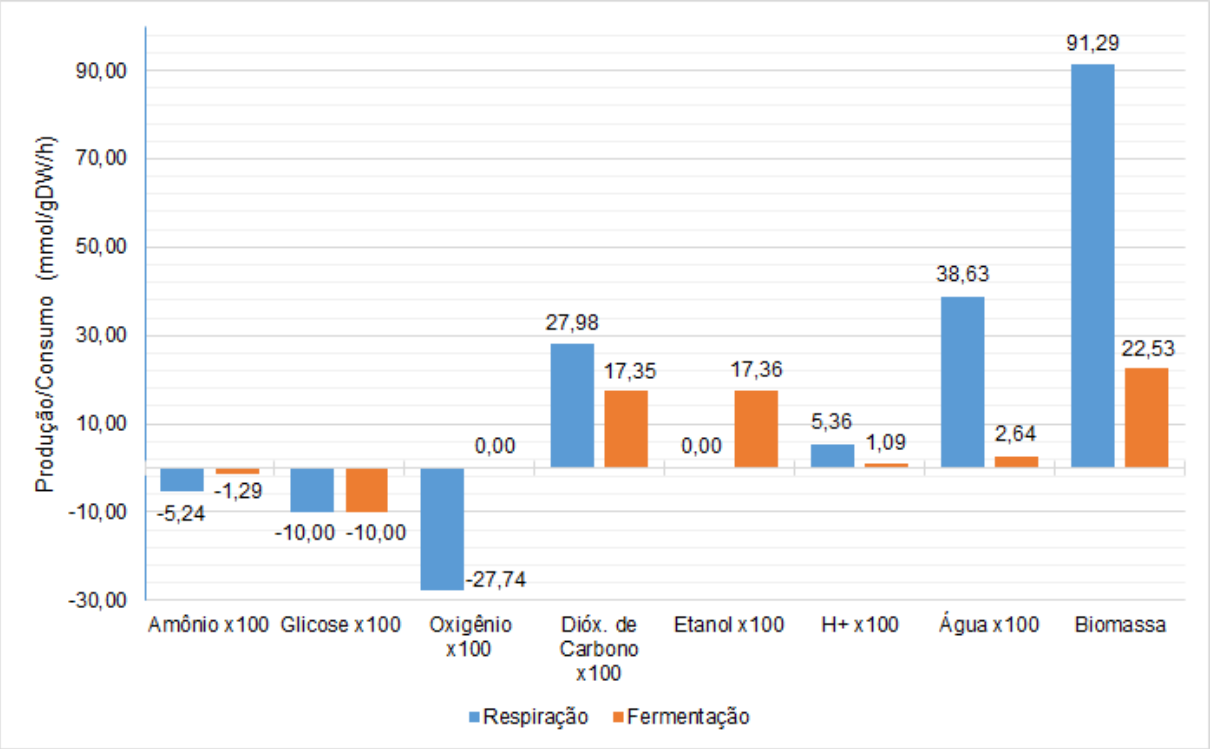


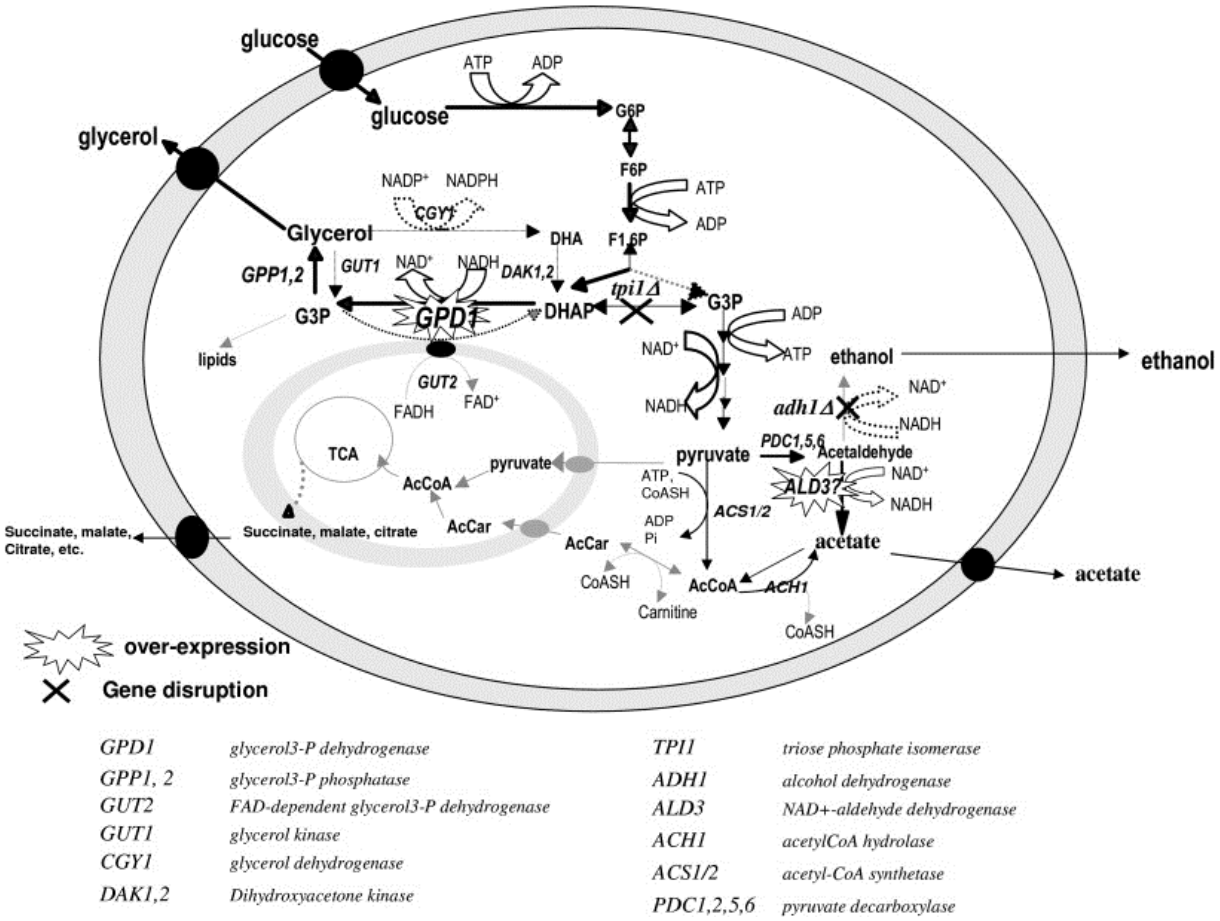
Figura 3.2: Fluxos de consumo (valores negativos) e produção (valores positivos) de alguns metabólitos na respiração e na fermentação.

Podemos notar que as simulações indicaram valores de fluxo compatíveis com a realidade, é observado, por exemplo, uma diminuição da biomassa e alta produção etanol no meio anaeróbio em relação ao meio aeróbio.

3.1.3 SIMULAÇÕES DA VIA DE PRODUÇÃO DE GLICEROL

Para testar a robustez do modelo e se ele conseguiria reproduzir experimentos realizados em laboratório, decidiu-se simular alguns experimentos de um estudo (CORDIER et al, 2007) onde se tenta elevar a produção de glicerol através de, entre outros artifícios, superexpressão do gene GPD1 (Glicerol-3-fosfato desidrogenase) e da deleção do gene ADH1 (Álcool desidrogenase) (Figura 3.3).

Para isso foi utilizado o *script addModels.m* para criar os modelos que simulam a deleção de ADH1 e a superexpressão de GPD1 usando as metodologias FBA e MOMA.



GPD1	*glycerol3-P dehydrogenase*	*TPI1*	*triose phosphate isomerase*
GPP1, 2	*glycerol3-P phosphatase*	*ADH1*	*alcohol dehydrogenase*
GUT2	*FAD-dependent glycerol3-P dehydrogenase*	*ALD3*	*NAD+-aldehyde dehydrogenase*
GUT1	*glycerol kinase*	*ACH1*	*acetylCoA hydrolase*
CGY1	*glycerol dehydrogenase*	*ACS1/2*	*acetyl-CoA synthetase*
DAK1,2	*Dihydroxyacetone kinase*	*PDC1,2,5,6*	*pyruvate decarboxylase*

Figura 3.3 (CORDIER et al., 2007): Estratégia de engenharia genética de *S. cerevisiae* utilizada pelo estudo de Cordier et al. (2007) para elevar a produção de glicerol e que foi simulada em parte pelo presente trabalho.

O estudo realizou algumas combinações de superexpressões e deleções a fim elevar a produção de glicerol. A seguir é apresentada a tabela 3.1 com as produções de etanol, glicerol e biomassa obtidas pelo estudo:

Tabela 3.1 (CORDIER et al., 2007): Produtos da fermentação de glicose a partir de diferentes linhagens geneticamente modificadas

	Produção (g / 100g de glicose)		
Modificação genética	**Biomassa**	**Etanol**	**Glicerol**
Selvagem	12±0,5	41±3,0	2,0±0,4
Superexpressão de GPD1	10±0,3	32±3,0	19,3±1,4
Superexpressão de GPD1 e deleção de TPI1	9±0,4	19±2,0	36±3,6
Deleção de ADH1	11±0,5	29±2,0	19±2,5
Superexpressão de GPD1 e de ALD3	10±1,3	33±1,0	18,5± 2,0
Superexpressão de GPD1 e de ALD3 e deleção de TPI1	9±1,4	19±2,0	42±2,5
Superexpressão de GPD1 e deleção de TPI1 e de ADH1	9.1±0,5	8±0,4	46±2.5
Superexpressão de GPD1 e de ALD3 e deleção de ADH1 e TPI1	8.1±1,3	8±1,4	46±2,7
Superexpressão de GPD1, de ALD3 e de FPS1 e deleção de ADH1 e TPI1	9±1,4	12±2,8	46±1,7

A primeira simulação consistiu na deleção de ADH1 do modelo selvagem. A seguir, a figura 3.4 apresenta um gráfico com as diferenças mais significativas encontradas entre a simulação do modelo selvagem e do mutante, simulado utilizando FBA e MOMA:

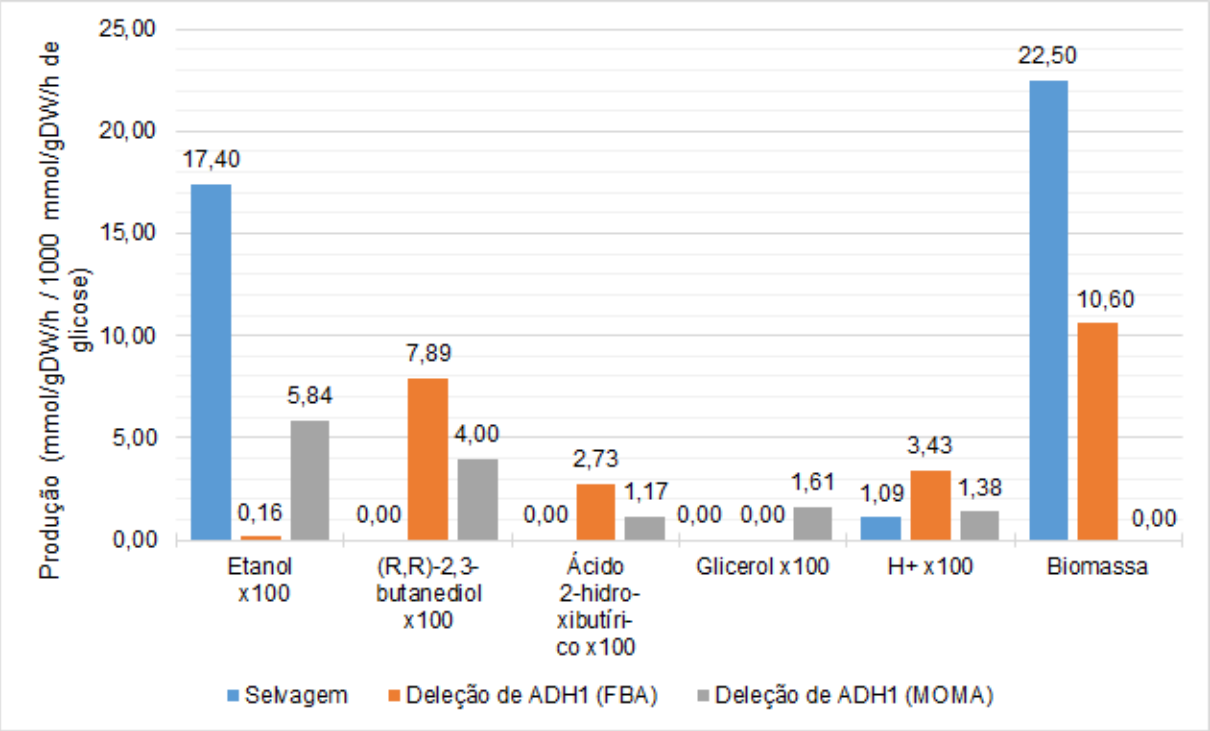


Figura 3.4: Principais produtos obtidos pelas simulações fermentação com deleção de ADH1.

Podemos notar que a simulação através do MOMA, embora tenha produzido glicerol, não gerou um resultado interessante para comparações, já que não há biomassa produzida, o que significa, biologicamente falando, que a levedura estaria morta. Este resultado pode ter ocorrido devido ao fato que o MOMA não se preocupa com a função objetivo, mas sim em buscar um conjunto de fluxos o mais próximo possível dos fluxos da linhagem selvagem. Sendo assim, foi decidido prosseguir os testes dessa etapa utilizando apenas o resultado da abordagem FBA, no qual podemos notar que a deleção de ADH1 gerou uma queda significativa na produção de etanol e também causou uma diminuição pela metade na produção de biomassa. Porém, não foi observado um aumento significativo de glicerol (podemos considerar que nada foi produzido), como era de se esperar. Além disso, existe uma produção inesperada de (R,R)-2,3-butanediol e de ácido 2-hidroxibutírico que ocorre na tentativa do sistema equilibrar o balanço redox e provavelmente está desviando o fluxo que deveria levar à produção de glicerol.

Na tentativa de direcionar a simulação com FBA para os resultados encontrados experimentalmente, a função objetivo foi alterada para contemplar a maximização da biomassa e o equilíbrio no balanço redox. Para tal foi inserido um segundo termo na função objetivo, f, deixando-a na forma $f = [Biomassa] - x\,[H+]$, que

reflete o desejo de maximizar a biomassa e minimizar a produção de H+. O fluxo de H+ na membrana reflete a disponibilidade de H+ interno que é gerado nas reações de oxi-redução: NAD → NADH + H+, dessa forma, alterar esse fluxo altera indiretamente as reações de oxi-redução. Como não sabíamos o peso da minimização na função objetivo, foi introduzida a variável x e esta foi estimada através de várias rodadas de simulação variando os valores de x de 0 até 1.

Assim foram testados vários valores do coeficiente x em busca da função objetivo que refletisse a realidade biológica. A seguir, a figura 3.5 apresenta um gráfico que exibe o valor da razão entre glicerol e etanol produzidos para cada simulação (variando o coeficiente de H+ na função objetivo), sendo que era buscada a configuração que apresentasse a razão mais próxima da razão entre glicerol e etanol obtida pela linhagem com deleção de ADH1 do experimento estudado. Deve ser observado que, para calcular a razão para os dados experimentais, foi necessário fazer conversão dos mesmos de gramas para mol, a fim de visualizar os dados na mesma unidade.

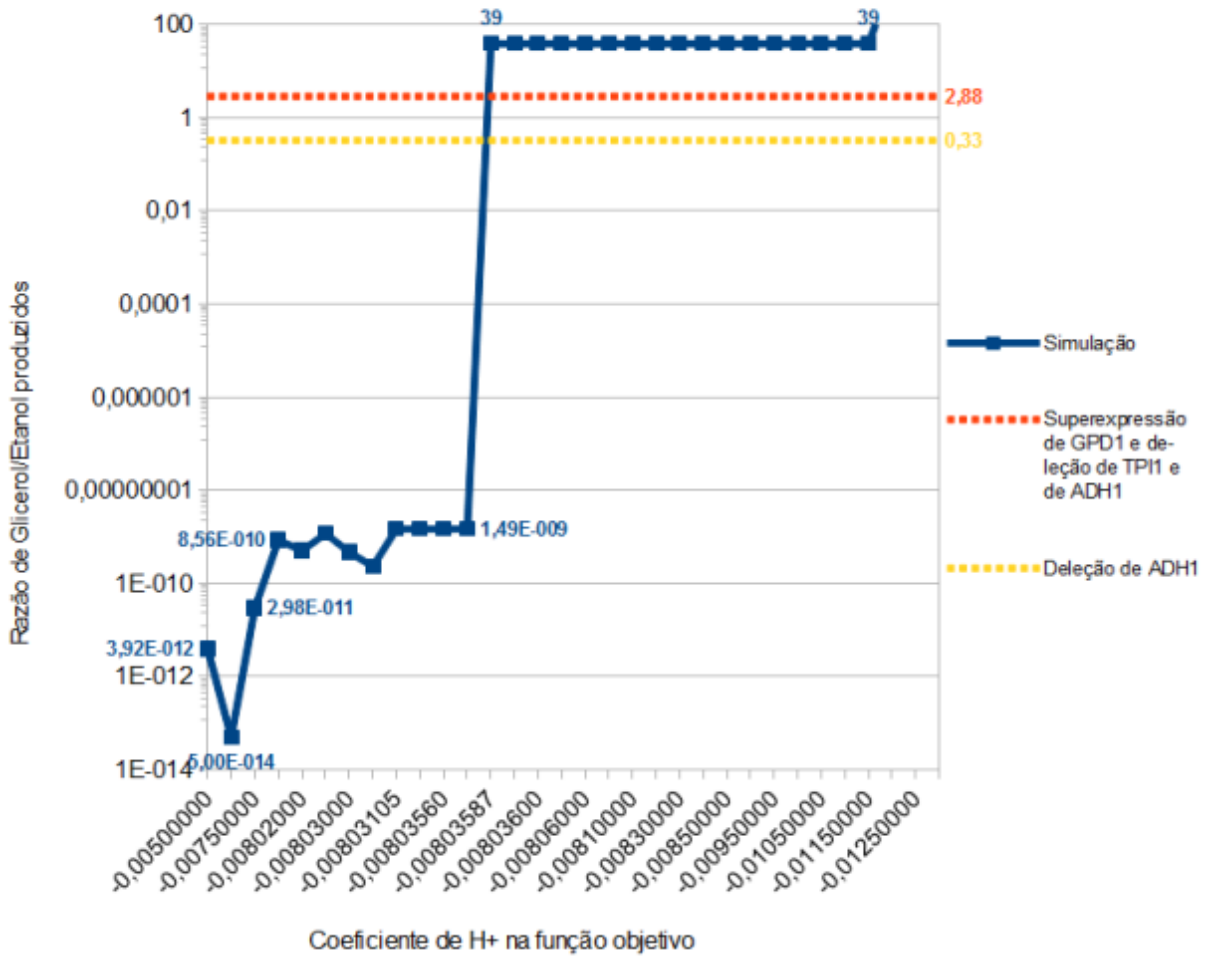


Figura 3.5: Gráfico em escala logarítmica apresentando a proporção entre glicerol e etanol produzidos na simulação em comparação com as proporções encontradas experimentalmente durante a deleção de ADH1 (amarelo) e superexpressão de GPD1 e deleção de TPI1 e ADH1 (vermelho) (valores do coeficiente de H+ não estão em escala).

O gráfico sugere que existe uma mudança no espaço de solução encontrado pela simulação quando o coeficiente do H+ na função objetivo está em torno de -0,008035. Essa mudança se mantém até valores em cerca de -0,012 para o coeficiente de H+. Assim, foi escolhido um valor intermediário (-0,009) para o coeficiente, a fim de dar continuidade às simulações. Deste modo, a figura 3.6 apresenta a seguir um gráfico com os produtos e o fluxo sobre as enzimas de interesse de uma simulação que utiliza como função objetivo *f = [Biomassa] – 0,009 [H+]* em comparação com a simulação que utiliza apenas a biomassa como função objetivo.

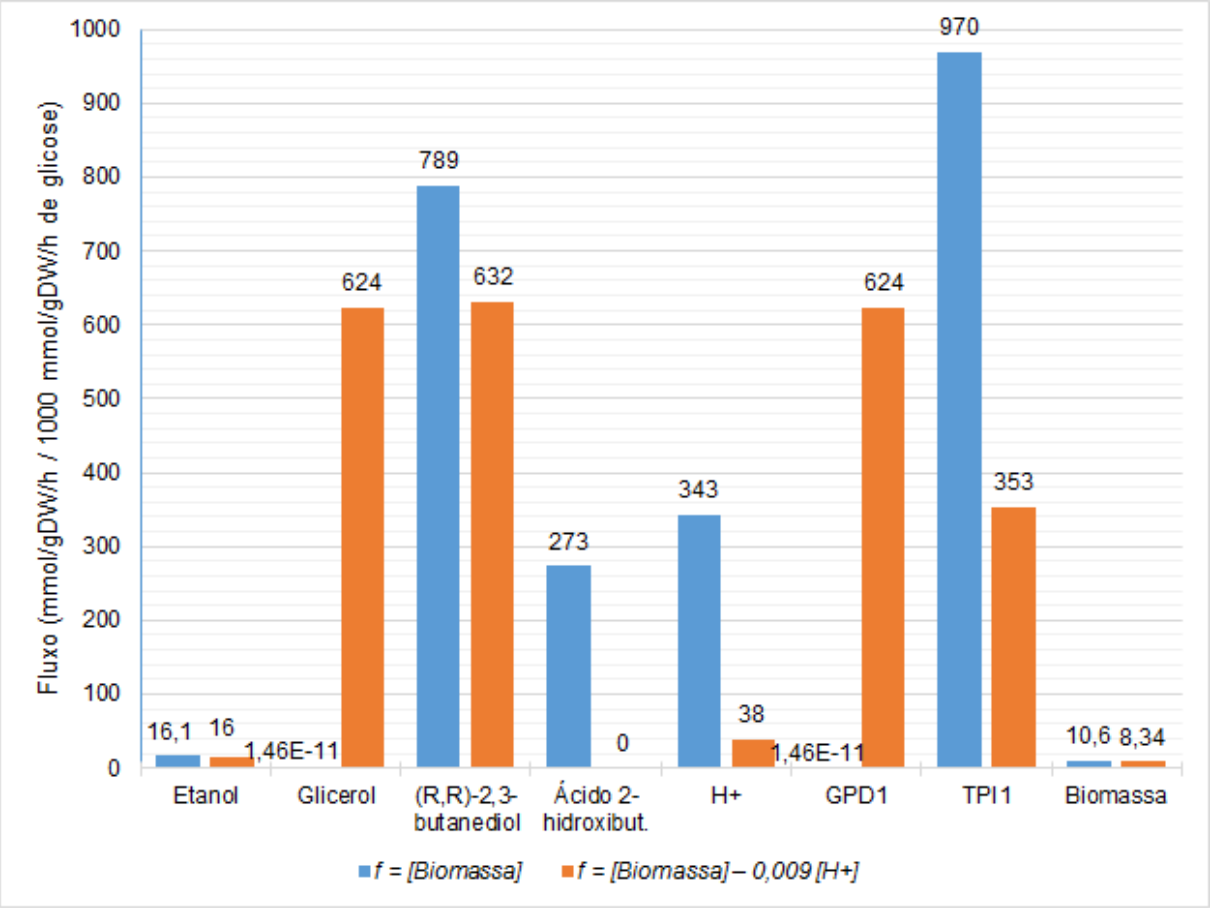


Figura 3.6: Comparação entre simulações com e sem a minimização da produção de H+ na função objetivo. Os primeiros 5 conjuntos de barras (de Etanol a H+) representam a produção dos respectivos metabólitos, já os 2 seguintes representam o fluxo sobre as enzimas GPD1e TPI1, respectivamente.

Pode-se verificar que, além de um significativo aumento na produção de glicerol, ocorre uma sutil diminuição na produção de (R,R)-2,3-butanediol e o fim da produção de ácido 2-hidroxibutírico, além de uma grande diminuição no acúmulo de H+. Também pode ser observado um aumento no fluxo da enzima GPD1, e uma queda no fluxo da enzima TPI1, que são, respectivamente, as duas alterações realizadas experimentalmente no genoma da levedura que junto com a deleção do ADH1 se mostraram mais eficientes para a produção de glicerol (tabela 3). Ou seja, a simulação automaticamente já indicou um caminho para aumento da produção de glicerol, já comprovado experimentalmente. Verifica-se assim o poder que as simulações têm de indicar tendências e possíveis alvos de modificação para determinados objetivos. Também mostra a importância dos *scripts* implementados, pois estes permitem a rápida construção de novos modelos e comparação global entre

eles, de forma a identificar qualquer mudança ocorrida no fluxo mesmo que estas ocorram em regiões distantes das alteradas propositalmente.

3.2 RESULTADOS E DISCUSSÕES DE SIMULAÇÕES ENVOLVENDO CONSUMO DE XILOSE

De forma a iniciar as simulações no modelo para fermentação utilizando-se xilose como fonte de carbono, primeiramente foi indicado no modelo uma alteração no limite de consumo de xilose indo de 0 para 1000 mmol/gDW/h, enquanto o limite de consumo de glicose permanece em zero. Na sequência, para cada simulação, foram inseridas no modelo *Yeast 5.0* as reações que simulam a inserção de genes necessários para a fermentação de xilose por *S. cerevisiae* e de genes candidatos a aumentar a eficiência do processo, conforme descrito abaixo.

Para a criação dos modelos foi usado o *script addModels.m* e para a comparação entre as diversas simulações foi usado o *script diffModels.m* com a opção de abordagem FBA, já que o MOMA é mais aplicável no caso de modelos com reações removidas em relação ao original, não sendo o caso destes testes.

3.2.1 FERMENTAÇÃO UTILIZANDO A VIA OXI-REDUTIVA

Na primeira simulação, as reações que representam a via oxi-redutiva foram incluídas no modelo; dessa forma, foi adicionada uma reação que converte xilose e NADPH em xilitol e NADP+, representando a enzima xilose redutase, e uma reação que converte xilitol e NAD+ em xilulose e NADH, representando a enzima xilitol desidrogenase. Desse modo, o esperado seria que a enzima xiluloquinase (endógena e já presente no modelo) passasse a ter mais xilulose disponível permitindo a produção de etanol através da via das pentoses fosfato, porém de maneira limitada, já que a transformação de xilitol em xilulose seria prejudicada devido ao desbalanço dos cofatores, lavando assim a um acúmulo do primeiro metabólito.

A seguir a figura 3.7 apresenta o gráfico com os destaques da comparação entre os fluxos da simulação de fermentação utilizando xilose e uma fermentação padrão utilizando glicose

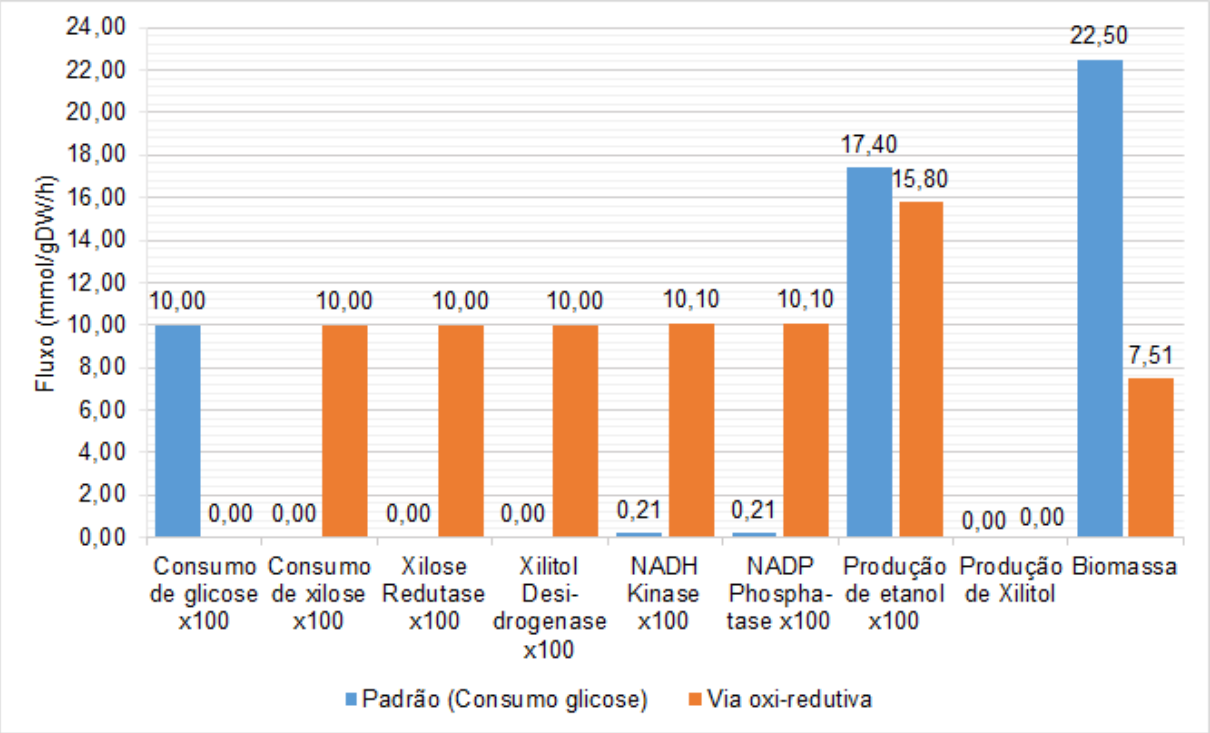


Figura 3.7: Comparação entre as simulações de fermentação utilizando glicose (Padrão) e consumindo xilose pela via oxi-redutiva.

Como é possível observar na figura 3.7, a simulação não mostrou o acúmulo de xilitol observado experimentalmente, sendo que o fluxo sobre as enzimas da via oxi-redutiva foi otimizado, levando a uma produção de etanol praticamente equivalente à observada pela fermentação de glicose. A causa de tal comportamento foi a presença de outras duas reações no modelo, cujos fluxos se alteraram durante a simulação de FBA e acabaram realizando o balanço de cofatores, mesmo que este não seja observado na realidade. Novamente, como ocorreu nas simulações do exemplo anterior (produção de glicerol), o modelo ajustou os fluxos globalmente alterando reações que não eram óbvias, que neste caso, foram responsáveis pelo ajuste natural do balanço redox. Como pode ser notada no gráfico, a análise comparativa das simulações levou à identificação de duas reações que tiveram uma elevação considerável em seus fluxos; são elas a NADH quinase, que transforma NADH e ATP em NADPH e ADP e a NADP fosfatase, que converte NADP+ e água em NAD+ e fosfato. O aumento do fluxo nessas reações leva a uma maior produção dos dois cofatores necessários para as reações de oxidação e redução de xilose.

Na tentativa de entender a função dessas duas enzimas endógenas de *S. cerevisiae*, foram realizadas diversas buscas na literatura. Sobre a enzima NAPH

quinase foi identificado um artigo (HOU et al., 2009) até então desconhecido pelo nosso grupo, que experimentalmente realiza a superexpressão desse gene em leveduras geneticamente modificadas para consumo de xilose utilizando a via oxi-redutiva e mostra que realmente existe uma melhora no balanço redox da primeira enzima da via, mas que esta melhora acaba acarretando no aumento da produção de xilitol, pois o desbalanço redox na segunda enzima da via continua ocorrendo. Este resultado é bem coerente com as nossas simulações, sendo que a função da segunda enzima que apareceu com uma elevação do fluxo (NADP fosfatase) é exatamente resolver o desbalanço redox da segunda enzima da via.

Durante as buscas por trabalhos da literatura sobre a enzima NADP fosfatase foi verificado que até o momento não foi identificado nenhuma sequência de gene que codifique para essa enzima em nenhum organismo (SPAANS et al., 2015). A atividade de NADP fosfatase tem sido detectada em vários organismos, mas o gene ainda não é conhecido. O único gene descrito até o momento com atividade de NADP fosfatase foi de *Methanococcus jannaschi* (KAWAI; MURATA, 2008) e não possui apenas essa atividade, sendo bifuncional NADP+ fosfatase/NAD+ quinase. Dessa forma, não foi possível ainda realizar experimentos que comprovem a solução teórica encontrada nesse trabalho.

3.2.2 FERMENTAÇÃO UTILIZANDO A VIA DE ISOMERIZAÇÃO

A fim de simular a fermentação de xilose através da via de isomerização, foram criados outros dois modelos, sendo o primeiro o modelo original com a adição de uma reação que converte xilose diretamente em xilulose, representando a enzima xilose isomerase e outro com a presença, além dessa enzima, das enzimas da via oxi-redutiva. Como não há cofatores envolvidos na reação da xilose isomerase, o esperado é que, na simulação, o organismo consuma toda a xilose disponível, apresentando uma boa produção de etanol.

O gráfico da figura 3.8 apresenta a comparação entre fluxos das simulações desses modelos e da simulação de uma fermentação padrão com glicose.

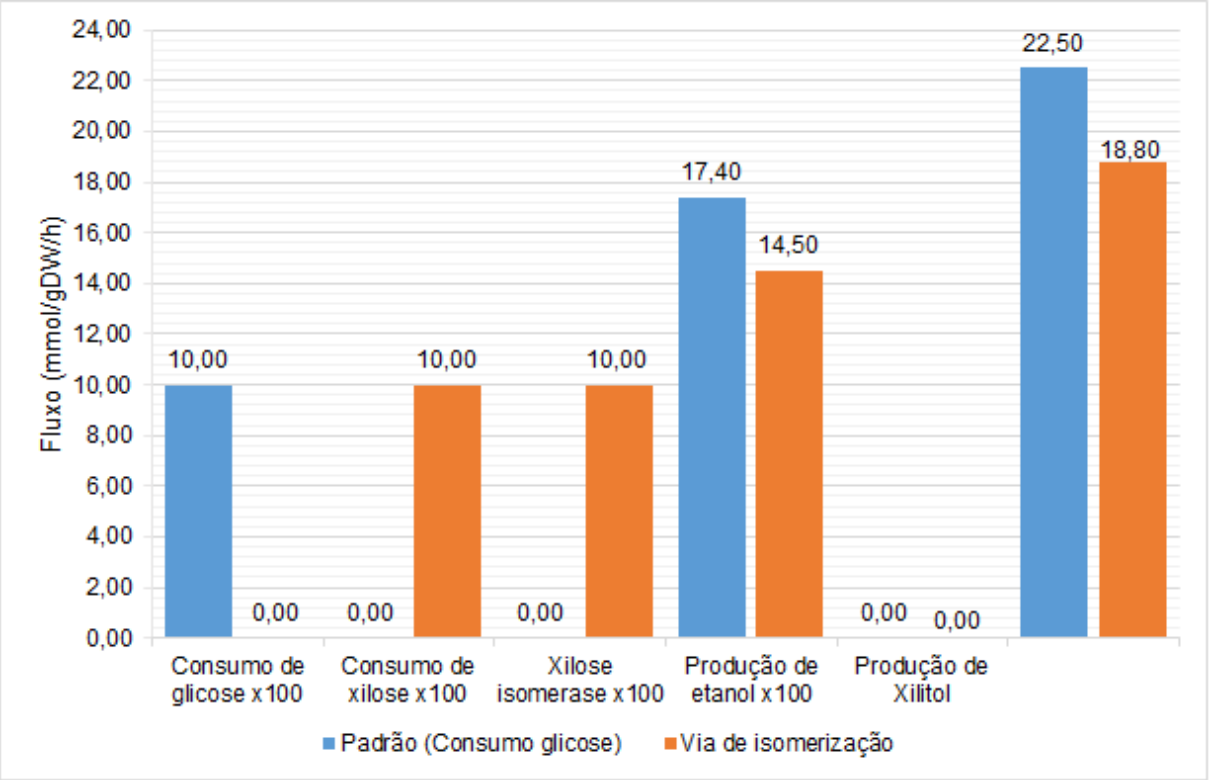


Figura 3.8: Comparação entre as simulações de fermentação utilizando glicose (Padrão) e consumindo xilose pela via de isomerização. Os valores dos fluxos nos modelos com e sem a via oxi-redutiva são idênticos, por esse motivo os valores são apresentados apenas uma vez.

Após a simulação notou-se não haver diferença entre as simulações com ou sem a via oxi-redutiva no caso da presença da xilose isomerase, isso já é esperado porque, como não há nenhuma dependência de cofatores para a reação da xilose isomerase, é natural que todo o fluxo de consumo passe por ela, já que existem gastos energéticos e consequente queda na produção de biomassa envolvidos no balanceamento dos cofatores na via oxi-redutiva. Assim, o sistema com a inclusão da xilose isomerase consome toda a xilose disponível e a converte em etanol, obtendo uma produção semelhante à observada no crescimento anaeróbico de glicose.

3.2.3 FERMENTAÇÕES UTILIZANDO AS VIAS DE ASSIMILAÇÃO DE XILOSE ASSOCIADAS À VIA DA FOSFOQUETOLASE

Por fim, foram criados modelos que simulam a inclusão da via da fosfoquetolase na levedura, trabalhando tanto com a via de isomerização quanto com a via oxi-redutiva. A via da fosfoquetolase é uma opção interessante, pois desvia o fluxo metabólico da xilose evitando passar (ou passando parcialmente) sobre a via da

pentose fosfato, que é conhecida por ser muito complexa e, portanto, limitante da velocidade da reação de conversão de xilose para etanol. Para a inclusão da via, foram adicionadas três novas reações, a primeira, representando a enzima fosfoquetolase (PK), que converte xilulose 5-fosfato e fosfato em acetil fosfato, gliceraldeído 3-fosfato e água, a segunda, representando a fosfotransacetilase (PTA), que converte acetil fosfato e a coenzima A em acetil-CoA e fosfato, e por último a enzima acetaldeído desidrogenase (ACDH) que é representada por uma reação que converte acetil-CoA e NADH em acetaldeído, coenzima A e NAD+. Além das reações, também foi adicionado o metabólito acetil fosfato, que não estava presente no modelo original.

Dado os resultados anteriores, o esperado nessa fase era que a inclusão da via da fosfoquetolase promovesse o balanceamento dos cofatores na via oxi-redutiva e que não tivesse efeito significativo sobre a via de isomerização. O gráfico da figura 3.9, apresenta a comparação entre os fluxos dos modelos com a via da fosfoquetolase e o modelo padrão.

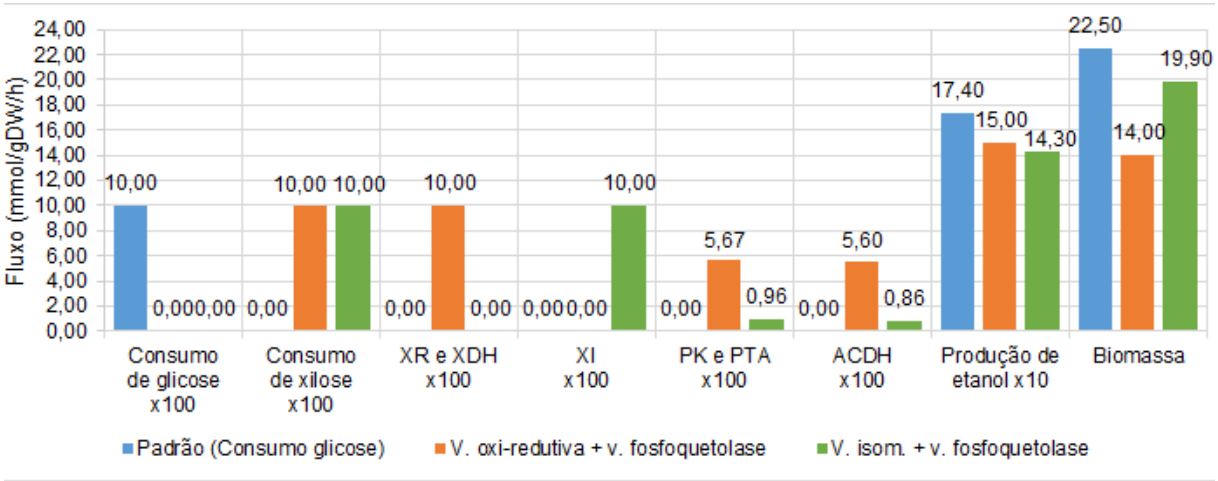


Figura 3.9: Comparação entre os fluxos de simulações das vias oxi-redutiva e de isomerização com a adição de via da fosfoquetolase e a simulação padrão. Note que os fluxos sobre as enzimas XR e XDH são iguais, por isso foram agrupados, assim como sobre as enzimas PK e PTA.

Podemos notar que, ao associar a via da fosfoquetolase com a via oxi-redutiva, ocorre um grande fluxo nas reações da via da fosfoquetolase, o que sugere que esta via está servindo como equilibrador dos cofatores da via oxi-redutiva. Já no modelo com a via de isomerização, o fluxo sobre a fosfoquetolase é bem menor, já que o

metabolismo da xilose pela via de isomerização não necessita de balanceamento de cofatores.

3.2.4 ANÁLISES COMPARATIVAS E CONCLUSÕES FINAIS

De forma a resumir e facilitar a discussão de todos as análises anteriores, está sendo apresentado na figura 3.10 um gráfico comparativo entre as fermentações de glicose e de xilose utilizando a via oxi-redutiva; a via de isomerização; a via oxi-redutiva e a via da fosfoquetolase; e a via xilose isomerase e a via da fosfoquetolase.

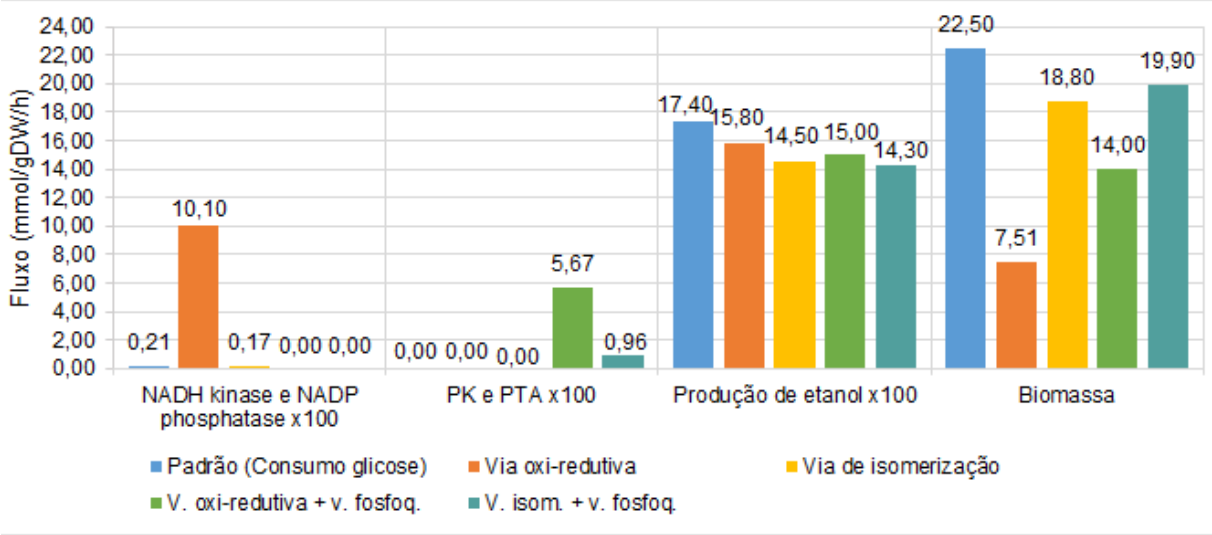


Figura 3.10: Comparação entre modelos modificados que consomem xilose e o modelo padrão consumindo glicose.

A partir desse gráfico, é possível enxergar algumas diferenças importantes entre cada simulação. Podemos notar que o modelo com a inclusão da via de isomerização apresenta uma vantagem sobre a via oxi-redutiva por não causar um desbalanço redox, minimizando, assim, o fluxo sobre as enzimas NADH quinase e NADP fosfatase que possuem um gasto de ATP que acaba afetando negativamente na formação da biomassa. Por outro lado, a maior produção de biomassa gerada pela via de isomerização causa uma produção de etanol ligeiramente menor em relação à via oxi-redutiva, isso se deve muito provavelmente ao fato da xilose isomerase não necessitar de cofatores, dispondo de mais energia para o crescimento e assim desviando o carbono do etanol para a formação da biomassa. Embora a via da isomerase produza um pouco menos etanol do que a via oxi-redutiva, essa via deve ser considerada como sendo mais interessante, pois a grande diferença encontrada

na formação da biomassa deve refletir num crescimento mais lento, prejudicando muito a eficiência da fermentação.

Outro fato interessante é notado quando a via da fosfoquetolase é associada à via oxi-redutiva. Nessa situação ocorre um grande fluxo sobre a via da fosfoquetolase e consequente ausência de fluxo nas enzimas NADH quinase e NADP fosfatase, ou seja, a via da fosfoquetolase passa a realizar o balanço de cofatores e o desvio sobre a via da pentose fosfato, possivelmente, permitindo a produção de etanol e forma mais eficiente. Esse fato também é notado quando em associação com via de isomerização, mesmo que essa não dependa dos cofatores para a fermentação, a via da fosfoquetolase também atua desviando o carbono da via da pentose fosfato. Por outro lado, podemos notar que quando a via da fosfoquetolase é inserida, tanto no modelo oxi-redutivo quanto no modelo de isomerização ocorre aumento da taxa de crescimento e consequente diminuição na produção de etanol. Isto ocorre devido à produção de acetato no fim da via da fosfoquetolase, pois a reação de formação de acetato gera fosfato, permitindo maior formação de ATP e consequente aumento na produção de biomassa, em detrimento de produção de etanol. De fato, nas simulações podem ser observadas produções de acetato somente nos modelos com a inclusão da via da fosfoquetolase, sendo produzido 3,1 mmol/gDW/h de acetato no modelo oxi-redutivo com fosfoquetolase e 4,43 mmol/gDW/h no modelo de isomerização com fosfoquetolase.

Por fim, é possível concluir que a fermentação através da via oxi-redutiva depende de alguma forma de balanço de cofatores, sendo que estudos sobre a viabilidade da superexpressão das enzimas NADH quinase e NADP fosfatase estão sendo realizados e podem apresentar uma boa alternativa no futuro, assim como a já conhecida via da fosfoquetolase. De forma geral, a via de isomerização se apresenta como uma alternativa bastante interessante já que não provoca desbalanço nos cofatores e não afeta a produção de biomassa. Porém as simulações realizadas nesse trabalho não conseguem mostrar o principal defeito observado experimentalmente para essa via, ou seja, a enzima da xilose isomerase não possui atividade catalítica quando expressa em *S. cerevisiae*, provavelmente devido a incorreto enovelamento dessa proteína, como já discutido no capítulo 1. Dessa forma, o capítulo 4 dessa tese foi dedicado ao uso de ferramentas computacionais de modelagem estrutural e métodos estatísticos de reconhecimento de padrões na tentativa de estudar o

problema com mais detalhes e assim identificar novas xilose isomerases funcionais em *S. cerevisiae.*

Capítulo 4

ANÁLISE FUNCIONAL E ESTRUTURAL DE XILOSE ISOMERASES

Diversos trabalhos anteriores avaliaram a funcionalidade de enzimas Xilose Isomerases (XI) originadas de outros organismos, principalmente em bactérias, expressas em *S. cerevisiae.* Sendo que alguns autores (TEUNISSEN; DE BONT, 2011; HA et al., 2011; MARIS et al., 2007; SHEN et al., 2012; BRAT et al., 2009; MADHAVAN et al., 2009; KUYPER et al., 2003; HECTOR, 2013; WALFRIDSSON et al., 1996) conseguiram mostrar experimentalmente exemplos de casos no qual a enzima teve atividade, mas o consenso na literatura é que a grande maioria dos casos de expressão heteróloga de XI em levedura resulta em inatividade da enzima (AMORE et al., 1989; TEUNISSEN; DE BONT, 2011; HA et al., 2011; BRAT et al., 2009; MOES et al., 1996; SARTHY et al., 1987; GÁRDONYI; HAHN-HÄGERDAL, 2003). O principal motivo da inatividade parece estar relacionado com o enovelamento incorreto da proteína dentro da célula eucariótica (GÁRDONYI E HAHN-HÄGERDAL, 2003). Dessa forma, o objetivo desta fase do estudo foi utilizar o conjunto de sequências de proteínas que mostraram funcionalidade em levedura e as que não se mostraram funcionais, listadas na tabela 4.1, para realizar análises de bioinformática utilizando ferramentas estatísticas que consideram a informação armazenadas na sequência linear e ferramentas de biologia estrutural para identificar novas xilose isomerases de outros organismos com potencial de serem funcionais em *S. cerevisiae*, além de ampliar o conhecimento dessas enzimas, avaliando as características estruturais presentes nas xilose isomerases que possam explicar sua funcionalidade na levedura.

Tabela 4.1: Xilose Isomerases encontradas na literatura que apresentam ou não funcionalidade em *S. cerevisiae*.

Organismo de origem	Atividade em *S. Cerevisiae*	Referência
Actinoplanes missouriensis	Não funcional	AMORE et al., 1989
Agrobacterium tumefaciens	Não funcional	BRAT et al., 2009
Arabdopsis thaliana	Não funcional	BRAT et al., 2009
Arthrobacter aurescens	Não funcional	TEUNISSEN; DE BONT, 2011
Bacillus licheniformis	Não funcional	BRAT et al., 2009
Bacillus subtilis	Não funcional	AMORE et al., 1989
Bifidobacterium longum	Não funcional	HA et al., 2011
Burkholderia phytofirmans	Não funcional	TEUNISSEN; DE BONT, 2011
Burkholderia xenovorans	Não funcional	BRAT et al., 2009
Clostridium thermosulfurogens	Não funcional	MOES et al., 1996
Escherichia coli	Não funcional	SARTHY et al., 1987
Haemophilus somnus	Não funcional	TEUNISSEN; DE BONT, 2011
Lactobacillus pentosus	Não funcional	TEUNISSEN; DE BONT, 2011
Leifsonia xyli	Não funcional	BRAT et al., 2009
Physcomitrella patens	Não funcional	TEUNISSEN; DE BONT, 2011
Robiginatalea biformata	Não funcional	BRAT et al., 2009
Saccharophagus degradans	Não funcional	BRAT et al., 2009
Salmonella typhimurium	Não funcional	BRAT et al., 2009
Staphylococcus xylosus	Não funcional	BRAT et al., 2009
Streptomyces diastaticus	Não funcional	BRAT et al., 2009
Streptomyces rubiginosus	Não funcional	GÁRDONYI; HAHN-HÄGERDAL, 2003
Thermatoga marítima	Não funcional	TEUNISSEN; DE BONT, 2011
Xantomonas campestris	Não funcional	BRAT et al., 2009
Bacteroides fragilis	Funcional	TEUNISSEN; DE BONT, 2011
Bacteroides stercoris	Funcional	HA et al., 2011
Bacteroides thetaiotaomicron	Funcional	MARIS et al., 2007
Ciona intestinalis	Funcional	TEUNISSEN; DE BONT, 2011
Clostridium difficile	Funcional	SHEN et al., 2012
Clostridium phytofermentans	Funcional	BRAT et al., 2009
Fusobacterium mortiferum	Funcional	TEUNISSEN; DE BONT, 2011
Orpinomyces sp.	Funcional	MADHAVAN et al., 2009
Piromyces sp. E2	Funcional	KUYPER et al., 2003
Prevotella ruminicola	Funcional	HECTOR, 2013
Thermus thermophilus	Funcional	WALFRIDSSON et al., 1996

4.1 PIPELINE PARA IDENTIFICAÇÃO DE POSSÍVEIS XILOSE ISOMERASES FUNCIONAIS EM *S. CEREVISIAE*

Para automatizar a busca por XIs com potencial de serem funcionais em levedura foi elaborado um pipeline utilizando-se as ferramentas HMMER e ClustalW.

Os dados de entrada do pipeline consistem em duas listas de proteínas (arquivos *multifasta*), o primeiro contendo proteínas funcionais em *S. cerevisiae* e o segundo contendo proteínas conhecidamente não funcionais em *S. cerevisiae*. O programa Clustal (LARKIN et al., 2007) na sua versão de linha de comando (ClustalW - versão 2.0.11), é utilizado para realizar alinhamentos múltiplos globais entre as sequências de cada lista.

A partir dos dois alinhamentos do ClustalW, um contendo as proteínas funcionais e outro contendo as não funcionais, o pacote HMMER (EDDY, 1998) é utilizado em sua versão 3.1b1 para buscar proteínas semelhantes a cada conjunto. Este pacote reúne vários programas que têm por objetivo realizar a análise de similaridade entre sequências proteicas através de métodos de inferência probabilística utilizando perfis modelos ocultos Markov. Para o pipeline, foram utilizados três programas do pacote, descritos a seguir.

- **hmmbuild:** Constrói um perfil modelo oculto de Markov (arquivo hmm) a partir de um alinhamento múltiplo de sequências. Um perfil HMM é um modelo matemático que representa a estrutura primária de uma família de proteínas (texto baseado http://www.biology.wustl.edu/gcg/hmmerbuild.html).
- **hmmcalibrate:** Calibra as estatísticas do arquivo hmm a fim de aumentar a sensitividade das buscas que utilizarem este perfil hmm como isca. O programa compara o perfil hmm original com um grande número de sequências geradas aleatoriamente e computa a pontuação (*raw score*) da comparação entre cada sequência e o perfil hmm, depois ajusta a distribuição dessas pontuações para estimar os parâmetros estatísticos de distribuição de valores extremos (*EVD*) que serão armazenados no perfil hmm e usados para o posterior cálculo de *E-value* de uma comparação (texto baseado em http://www.biology.wustl.edu/gcg/hmmercalibrate.html).
- **hmmsearch:** Utiliza um perfil hmm para buscar em um banco de dados de sequências aquelas que são relacionadas com o conjunto de sequências que

originou o perfil (texto baseado em http://www.biology.wustl.edu/gcg/hmmersearch.html).

Dessa forma, dois modelos ocultos de Markov, um para cada lista, são construídos utilizando-se *hmmbuild* e calibrados com *hmmcalibrate*. Em seguida, é utilizado *hmmsearch* para buscar proteínas no banco de dados público do NCBI (NR/NCBI) na tentativa de trazer novas proteínas que tenham padrões de sequência que alinham bem com os padrões de sequências geradas pelos modelos de Markov, sendo que de cada modelo obtém-se uma lista com as proteínas encontradas, com seus respectivos *scores* e *E-values*.

A fim de encontrar e classificar proteínas existentes nos bancos de dados públicos que tenham padrões relacionados aos modelos gerados foi necessário construir um *script* utilizando a linguagem Perl, ao qual foi dado o nome de *hmmer_venn.pl*. Este *script* tem por função selecionar as sequências obtidas pelo *hmmsearch* que obedeçam a um *E-value* de corte (no caso deste estudo, menor que 1e-300) e separá-las em três categorias: potencialmente funcionais (presentes apenas na lista de proteínas gerada pelo modelo funcional), potencialmente não funcionais (presentes apenas na lista gerada pelo modelo não funcional) e intersecção (presentes nas listas dos dois modelos). Assim, a intersecção é descartada e, a partir das listas curadas de proteínas potencialmente funcionais e potencialmente não-funcionais, se dá início a uma nova rodada do pipeline, alinhando as proteínas de cada lista entre si através do ClustalW e assim por diante.

A ideia de serem realizadas rodadas de busca surge do desejo de se obter um modelo que não seja viciado em apenas poucas sequências iniciais, abrindo a oportunidade de se encontrar proteínas que não sejam extremamente semelhantes com as sequências iniciais, mas que ainda apresentem regiões conservadas que possam significar sua funcionalidade em *S. cerevisiae*. Porém, após serem realizados testes com várias rodadas percebeu-se que as proteínas obtidas começavam a se distanciar muito dos conjuntos iniciais, com, por exemplo, proteínas que incialmente tinham semelhança com o grupo de funcionais sendo classificada como não funcional, assim para a continuidade dos estudos decidiu-se utilizar três rodadas de iteração, já que nesse ponto eram observadas proteínas que atendiam ao equilíbrio desejado.

Ao finalizar as rodadas, é realizada a segunda fase do pipeline, que tem por objetivo a identificação nas sequências obtidas na primeira fase do maior número possível de motivos descritos pela patente *Xylose isomerase genes and their use in*

fermentation of pentose sugars (TEUNISSEN e DE BONT, 2011) como indicadores de funcionalidade de uma xilose isomerase em células eucarióticas. A patente descreve oito motivos, presentes em posições específicas da proteína e que variam entre simples resíduos quanto sequências de até seis aminoácidos, informados na tabela 4.2.

Tabela 4.2: Motivos descritos pela patente de Teunissen e De Bont (2011) a serem buscados nas xilose isomerases.

Sequência	**Abreviação**	**Posição**
Treonina – Glicina – Isoleucina – Lisina – Leucina – Leucina	TGIKLL	134-139
Treonina – Leucina – Alanina – Glicina – Histidina	TLAGH	274-278
Arginina – Tirosina – Alanina – Serina – Fenilalanina	RYASF	387-391
Metionina	M	91
Fenilalanina	F	230
Fenilalanina – Lisina	FK	264-265
Glicina	G	394
Alanina	A	431

Além da busca pelos motivos, decidiu-se incluir no pipeline uma análise de identidade das proteínas encontradas em relação a todas xilose isomerases já patenteadas (SHEN et al., 2012; CAIMI et al., 2013; DOBSON; KRUCKEBERG, 2013; GE et al., 2013; HITZ et al., 2013; HUGHES; BUTT, 2010; JORDAN et al., 2014; OTERO et al., 2002; RAJGARHIA et al., 2013; SANNY; STARK, 2011; SATOSHI et al., 2013; SUBBIAN et al., 2011; SUBBIAN et al., 2013; WINKLER et al., 2012), a fim que este pudesse ter uma segunda aplicação: o uso em indústrias de biotecnologia. Sendo assim, foi adicionado ao pipeline um filtro para seleção de xilose isomerases que apresentassem somente identidades menores que 70% em relação a qualquer xilose isomerase patenteada.

Os *scripts clustalW_P2P.pl* e *clustalW_identidades.pl* foram desenvolvidos a fim de automatizar a tarefa de cálculo de identidades. O primeiro recebe como parâmetros dois arquivos *multifasta* – no caso deste projeto, um com as proteínas encontradas pelo último *hmmsearch* da primeira fase e outro contendo as sequências patenteadas (foram obtidas 368 sequências patenteadas no site http://www.lens.org/)

– e realiza um alinhamento global utilizando ClustalW para cada par de sequências, produzindo como saída uma pasta contendo um arquivo “*.aln*” para cada ClustalW realizado. Já o *script clustalW_identidades.pl* recebe a pasta que contém os arquivos “*.aln*” obtidos pelo *script* anterior e cria uma tabela onde cada linha contém uma proteína da lista gerada pelo *hmmsearch* e os valores máximo e mínimo de identidade com as sequências patenteadas.

O *script verificaMotivos.pl* foi desenvolvido para verificar se cada uma das proteínas contidas em um *multifasta* passado como parâmetro contém os motivos descritos em um arquivo de apoio também passado como parâmetro (Figura 4.1). Este arquivo contém em cada linha a sequência de um motivo, a posição de início deste, e uma marcação binária indicando se o *script* tem deve procurar pelo motivo em qualquer posição da sequência. Isso permite a correção de possíveis deslocamentos na sequência da proteína, pois o *script* armazena o valor de deslocamento entre a posição indicada e a posição onde o motivo foi encontrado e verifica se os outros motivos estão, não só na posição indicada, mas também na posição levando em conta o deslocamento. Como saída, o *script* cria uma tabela com cada proteína contida no *multifasta* e a posição onde foram encontrados cada motivo (ou 0 no caso de motivo não encontrado).

No caso deste projeto, foram escolhidos os 3 motivos maiores (TGIKLL, TLAGH e RYASF) para serem procurados em qualquer lugar da proteína, assim os outros seriam procurados nas posições originais ou relativas às posições dos motivos maiores. Por exemplo, se em determinada sequência o motivo TGIKLL for encontrado na posição 135, e não na posição 134, o motivo TLAGH for encontrado na posição 275 e o motivo RYASF for encontrado na posição 389; a Metionina será procurada, além da posição 91, também nas posições 92 e 93, e assim também será feito para os outros motivos nesta sequência.

```
1  TGIKLL;134;1
2  TLAGH;274;1
3  RYASF;387;1
4  M;91;0
5  F;230;0
6  FK;264;0
7  G;394;0
8  A;431;0
```

Figura 4.1: Arquivo utilizado pelo *script verificaMotivos.pl* com os motivos a serem procurados em uma sequência. Para os três primeiros (terceira coluna igual a 1) o *script* buscará em qualquer posição da sequência e caso encontrado registrará o deslocamento em relação à posição original (segunda coluna). Para os cinco últimos, o *script* buscará o motivo na posição indicada no arquivo, e também na posição indicada somada a cada deslocamento registrado.

A figura 4.2 apresenta o esquema do pipeline como um todo.

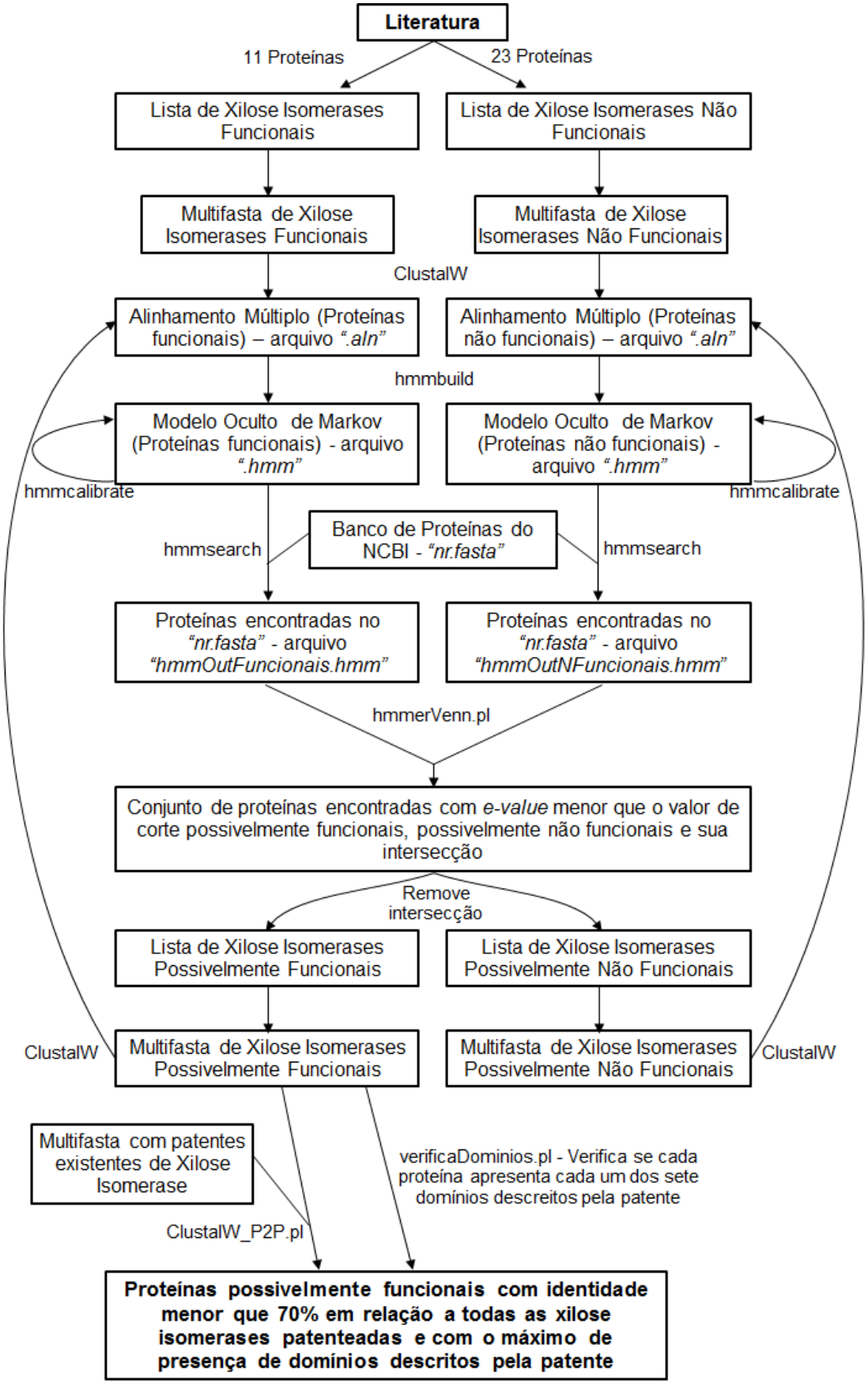


Figura 4.2: Esquema do pipeline desenvolvido.

Após as três iterações do processo de busca, pôde-se observar que nenhuma proteína se mostrou presente em ambas as listas, como é mostrado na figura 4.3. Assim, as 213 proteínas presentes na última lista de xilose isomerases potencialmente funcionais foram avaliadas quanto à presença dos motivos e 96 destas mostraram uma identidade menor que 70% em relação a todas as proteínas atualmente presentes em patentes.

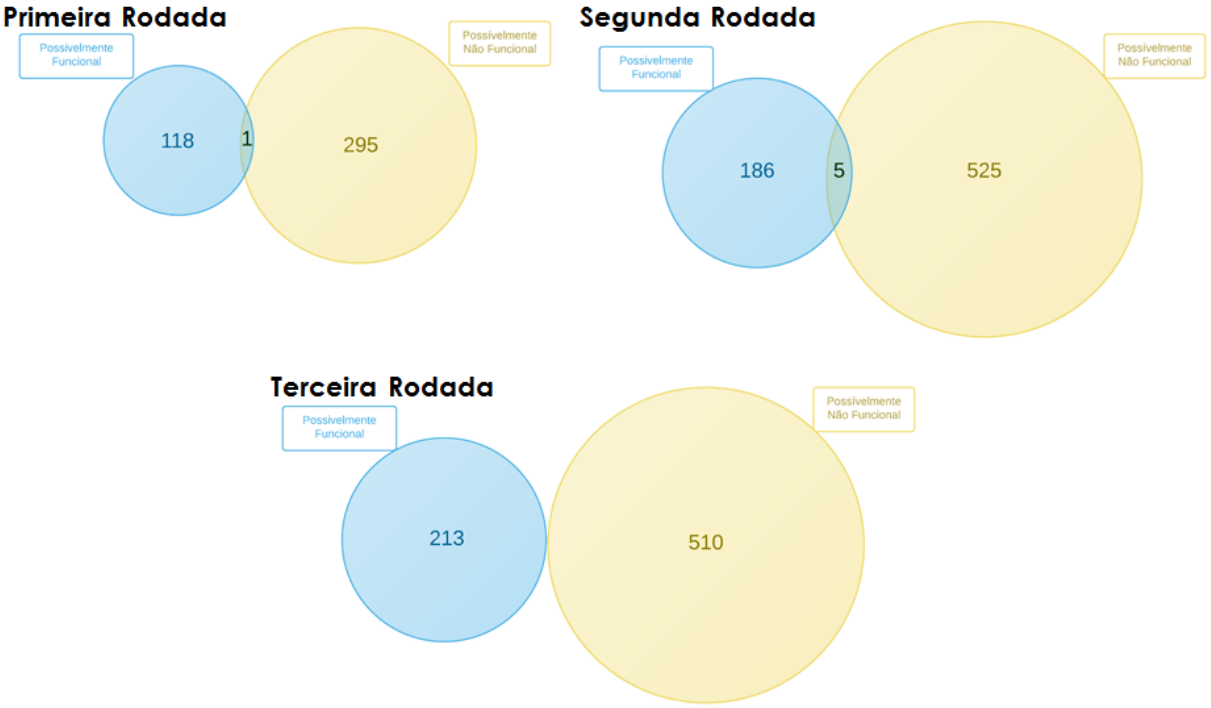


Figura 4.3: Diagramas de Venn representando os resultados das iterações 1, 2 e 3. Em azul são exibidas as quantidades de proteínas potencialmente funcionais encontradas e em amarelo as possivelmente não funcionais.

Na figura 4.4, é apresentado um histograma com a presença de motivos em cada grupo de proteínas: as funcionais e não funcionais da lista inicial, e as potencialmente funcionais e não funcionais encontradas pelo pipeline. Podemos notar que a grande maioria das proteínas não funcionais apresentam poucos motivos (menos que 4), isso também é observado nas sequências obtidas pelo pipeline. Já para as funcionais, grande parte possui mais de 5 motivos, sendo que 7 das 11 da literatura apresentam todos. Além disso, 30% das sequências encontradas pelo pipeline têm todos motivos, e apenas 25% menos que 4.

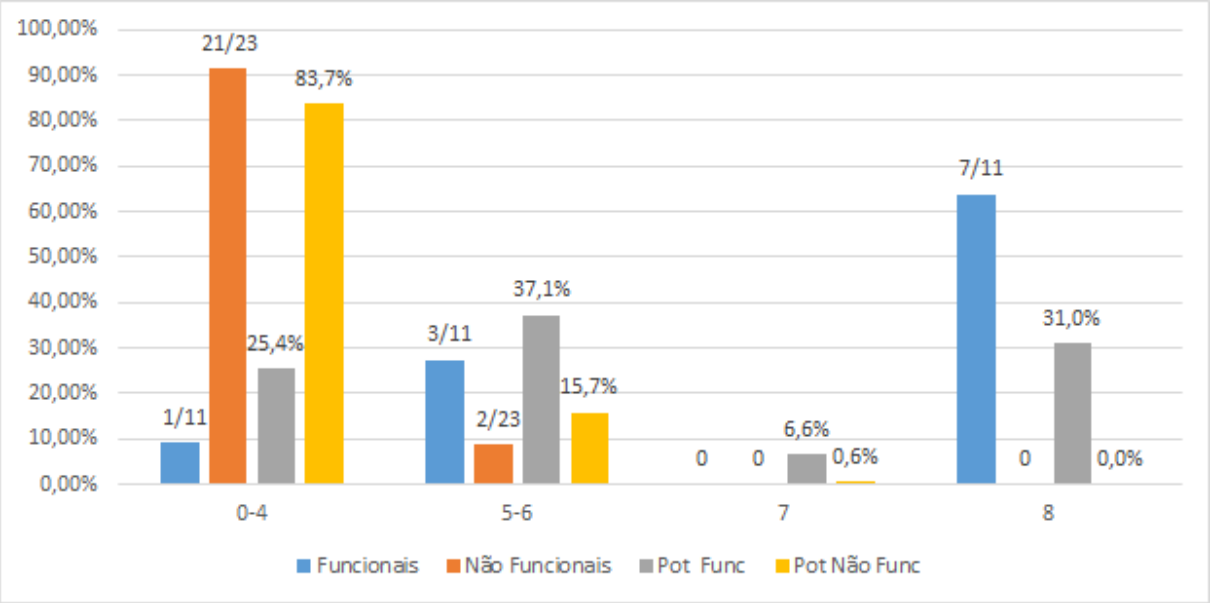


Figura 4.4: Histograma com a presença dos motivos em cada grupo de proteínas: funcionais da literatura, não funcionais da literatura, potencialmente funcionais e potencialmente não funcionais.

Algumas análises foram realizadas no sentido de verificar a possível relevância de cada um dos oito motivos para a funcionalidade da proteína em *S. cerevisiae*, ou seja, quais motivos se mostram presentes nas xilose isomerases funcionais e ausentes nas não funcionais e também nas sequências encontradas pelo pipeline. A tabela 4.3 compara a porcentagem de proteínas em cada um dos quatro grupos que possuem cada motivo de interesse.

Tabela 4.3: Porcentagem de proteínas de cada grupo que apresentam cada motivo. As duas primeiras linhas se referem às sequências potencialmente funcionais e não funcionais obtidas pelo pipeline e as duas últimas se referem às sequências já conhecidas como sendo funcionais ou não em *S. cerevisiae.*

	Motivo							
	TGIKLL na posição 134	**TLAGH na posição 274**	**RYASF na posição 387**	**M na posição 91**	**F na posição 230**	**FK na posição 264**	**G na posição 394**	**A na posição 431**
XIs Pot. Funcionais	40,38%	81,22%	76,53%	58,22%	93,90%	76,06%	88,26%	38,03%
XIs Pot. Não Funcionais	0,99%	96,64%	1,38%	2,77%	26,68%	28,06%	29,25%	15,22%
XIs Funcionais (Lista Inicial)	81,82%	90,91%	72,73%	90,91%	90,91%	90,91%	90,91%	63,64%
XIs Não Funcionais (Lista Inicial)	8,70%	39,13%	13,04%	4,35%	30,43%	30,43%	43,48%	8,70%

Ao observar a tabela, é possível notar que, como esperado, os motivos estão presentes na grande maioria das xilose isomerases funcionais, e em poucas não funcionais, sendo que isso também é observado, de maneira geral, para as xilose isomerases classificadas pelo pipeline. Outros detalhes também podem ser observados, como, apesar o motivo TGIKLL estar bastante presente nas proteínas funcionais (81,82%), ele aparece em relativamente poucas sequências do grupo potencialmente funcional (40,38%), já o motivo TLAGH é um dos motivos mais encontrado nas xiloses não funcionais (39,13%) e também tem uma aparição surpreendentemente alta nas potencialmente não funcionais (96,64%).

Este pipeline deu origem ao software XIMMER – Software para identificação de xilose isomerase, que foi registrado pela Agência de Inovação da Unicamp (INOVA) e atualmente está licenciado para a empresa Biocelere Agroindustrial Ltda. O módulo adicional de identificação de proteínas que fogem de patentes foi um dos fatores predominantes de interesse pela empresa, uma vez que viabiliza uma aplicação

industrial dessas enzimas, que caso apresentem alguma funcionalidade, estão livres para serem utilizadas comercialmente e passíveis de proteção.

4.2 RESULTADOS DA MODELAGEM E ANÁLISE ESTRUTURAL DE XILOSE ISOMERASES

Através do software YASARA as proteínas funcionais e não funcionais descritas na literatura foram submetidas ao processo de modelagem por homologia a fim de buscar características estruturais que possam auxiliar no entendimento de sua funcionalidade ou não em *S. cerevisiae*. Para tal, foram computadas as energias internas de cada proteína e seus monômeros, as energias de ligação entre monômeros e as energias de ligação entre dímeros, com o objetivo de encontrar alguma evidência que separasse o grupo funcional do não funcional. Também foram verificados a localização dos motivos funcionais descritos na patente, em especial na proteína originária da *Piromyces sp.*, reconhecida por ter um bom desempenho na *S. cerevisiae*.

A fim de se automatizar a construção dos modelos tridimensionais das xilose isomerases, foi adaptado para a linguagem *Python*, o *script hm_build*, fornecido pelo YASARA na linguagem *Yanaconda*. O parâmetro de entrada do *script* deve ser um arquivo fasta contendo a sequência de aminoácidos da proteína a ser modelada. Além disso, dentro do *script* é possível modificar diversos parâmetros utilizados pela macro *ExperimentHomologyModeling.* Os parâmetros e os valores utilizados na modelagem estão descritos a seguir.

- **psiblasts**: Número de iterações utilizadas pelo PSI-BLAST para alinhamento com os moldes do PDB. Foram utilizadas 3 iterações.
- **evalue**: Valor máximo do *e-value* dos alinhamentos a ser permitido. Foi utilizado o valor 0.5.
- **templates**: Número máximo de moldes utilizados para construção do modelo. Foi definido o uso de no máximo 5 moldes.
- **alignments**: Número máximo de alinhamentos ambíguos a serem considerados por molde. Foi definido considerar-se no máximo 5 alinhamentos ambíguos por molde.

- **oligostate**: Estado máximo de oligomerização, sendo suportado no máximo tetrâmeros (4). Foi escolhido o valor 4, uma vez que as xilose isomerases presentes no PDB são tetraméricas.
- **termextension**: Número máximo de resíduos de laços não alinhados a serem inseridos na região terminal. Foi definido o valor 0.

Dado que a xilose isomerase consiste em um tetrâmero, o método *ExperimentHomologyModeling* gera um modelo contendo quatro monômeros, nomeados de 'A' a 'D'. Porém, para cada proteína modelada, essa nomeação não segue um padrão, ou seja, as interfaces de ligação entre os monômeros não são necessariamente as mesmas em duas proteínas diferentes. Por exemplo, em determinada proteína, a interface de contato entre os monômeros A e B é equivalente à interface de contato entre os monômeros A e C de outra proteína. Isso faz com que as energias de ligação que serão calculadas posteriormente não serão equivalentes entre as proteínas (energia de ligação entre A e B de uma proteína será equivalente à energia de ligação entre A e C de outra).

Sendo assim, se faz necessário uma normalização dos nomes dos monômeros das proteínas. Para isso foi desenvolvido o *script comparaDimeros.py*, onde os nomes dos monômeros serão dados seguindo os nomes dados aos monômeros de uma mesma proteína “padrão” (no caso, foram utilizados os nomes dos monômeros do modelo da xilose isomerase de *Piromyces sp*.).

O *script* consiste em realizar alinhamentos estruturais, utilizando a função *AlignMol*, entre cada dímero de cada proteína modelada e cada dímero da proteína escolhida arbitrariamente como padrão (neste caso, *Piromyces sp.*). Após o alinhamento, o script compara a quantidade de resíduos alinhados em cada alinhamento, permitindo assim mapear os dímeros de uma proteína em outra.

Figura 4.5: Exemplo de funcionamento do *script comparaDimeros.py*. Utilizando a proteína 1 como modelo, os nomes de cada monômero da proteína 2 são trocados a fim de representarem as mesmas regiões de contato da proteína padrão. O processo é repetido para todas as proteínas modeladas.

Após a padronização dos nomes dos monômeros, foi desenvolvido o *script bind_energy.py*, que automatiza os experimentos de cálculo das energias internas e de ligação. Sendo assim, o *script* realiza um experimento de minimização de energia da proteína e em seguida coleta os valores das energias internas da proteína inteira e de cada monômero; depois, calcula a energia de ligação entre cada monômero e os restantes, e entre cada par de dímeros.

As proteínas provenientes das espécies *Burkholderia xenovorans*, *Streptomyces rubiginosus*, *Streptomyces diastaticus*, *Bifidobacterium longum*, *Burkholderia phytofirmans*, *Arthrobacter aurescens*, *Actinoplanes missouriensis* e *Staphylococcus xylosus,* todas não funcionais, apresentaram alguns valores inesperados de energias internas (valores positivos sendo que deveriam ser negativos) e/ou de ligação (valores negativos sendo que deveriam ser positivos), indicando uma provável modelagem incorreta, sendo excluídas das análises posteriores. A seguir são apresentados gráficos das energias de ligação e das energias internas de cada monômero, que já são calculados no momento da modelagem.

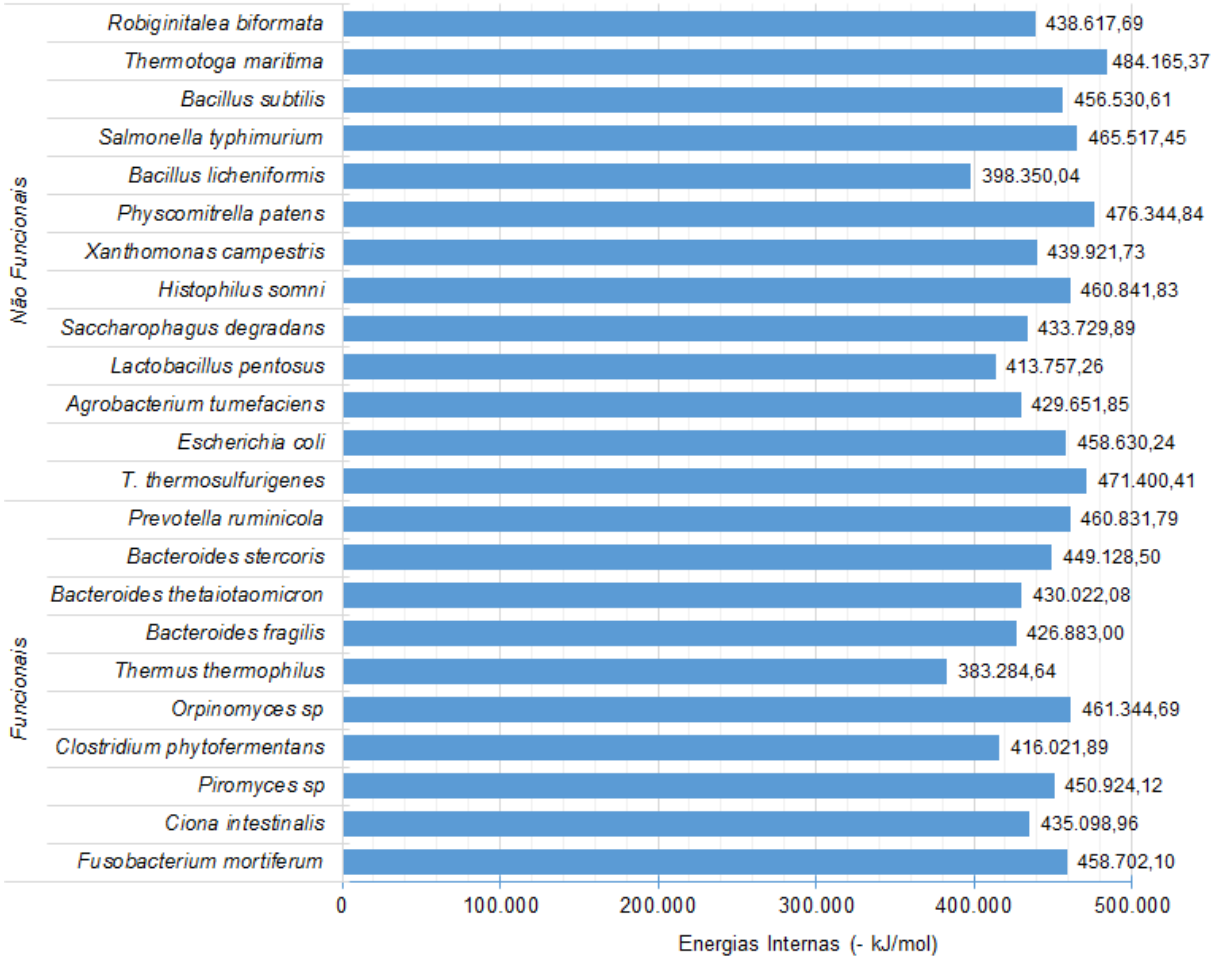


Figura 4.6: Gráfico das energias internas de cada xilose isomerase modelada após a minimização de energia.

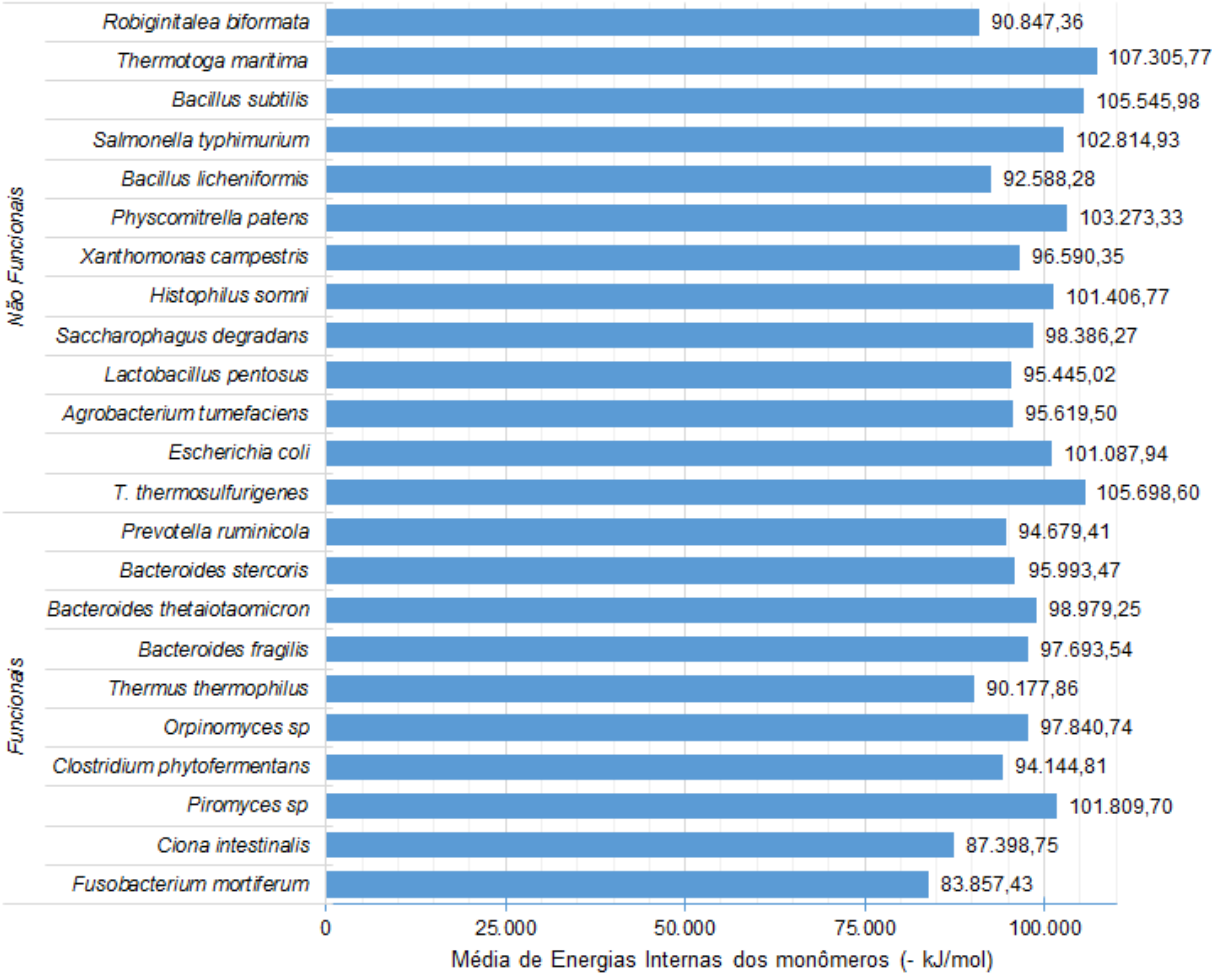


Figura 4.7: Gráfico das médias das energias internas de cada xilose isomerase modelada após a minimização de energia.

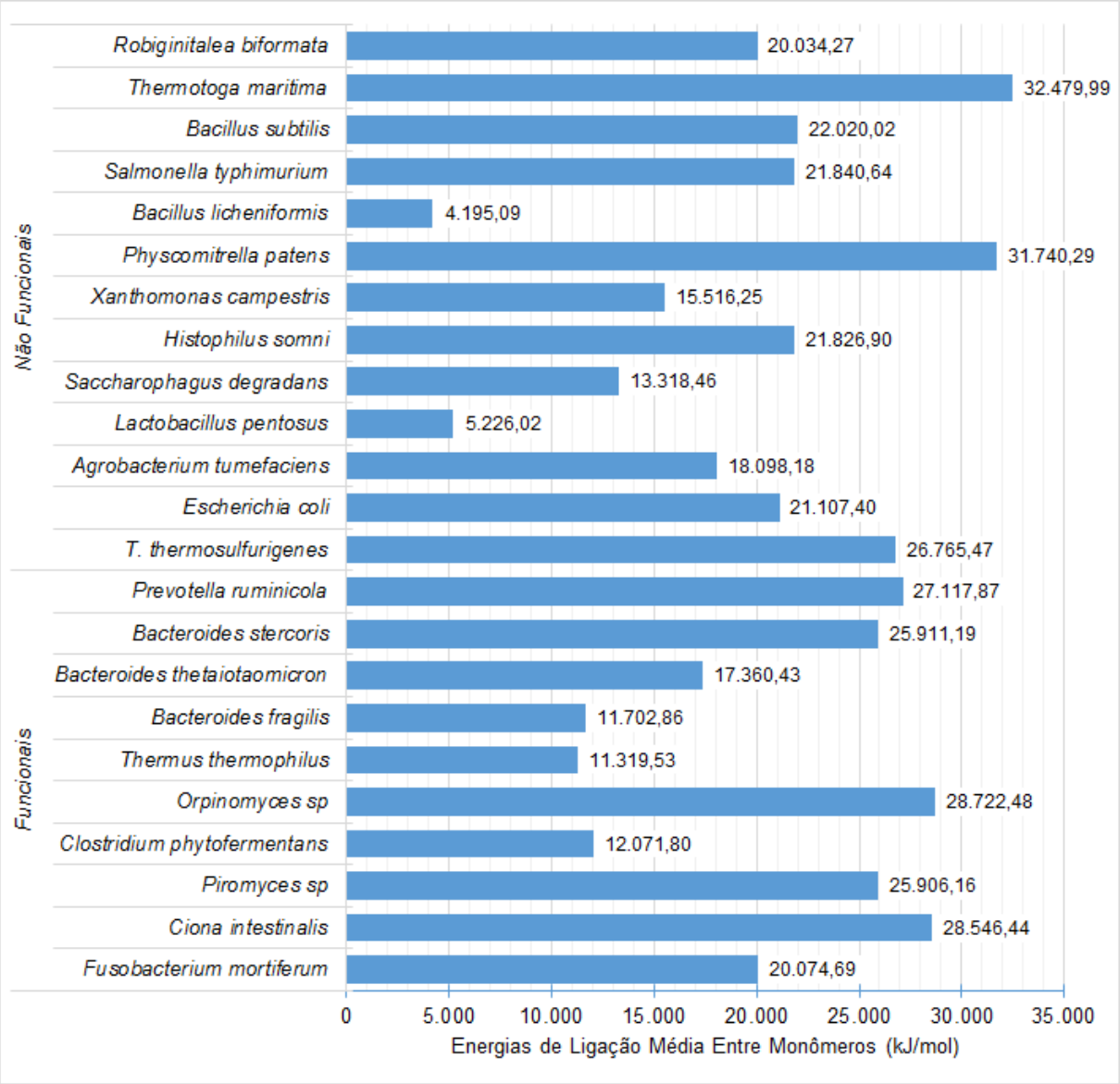


Figura 4.8: Gráfico das médias das energias de ligação entre os monômeros de cada xilose isomerase modelada após a minimização de energia.

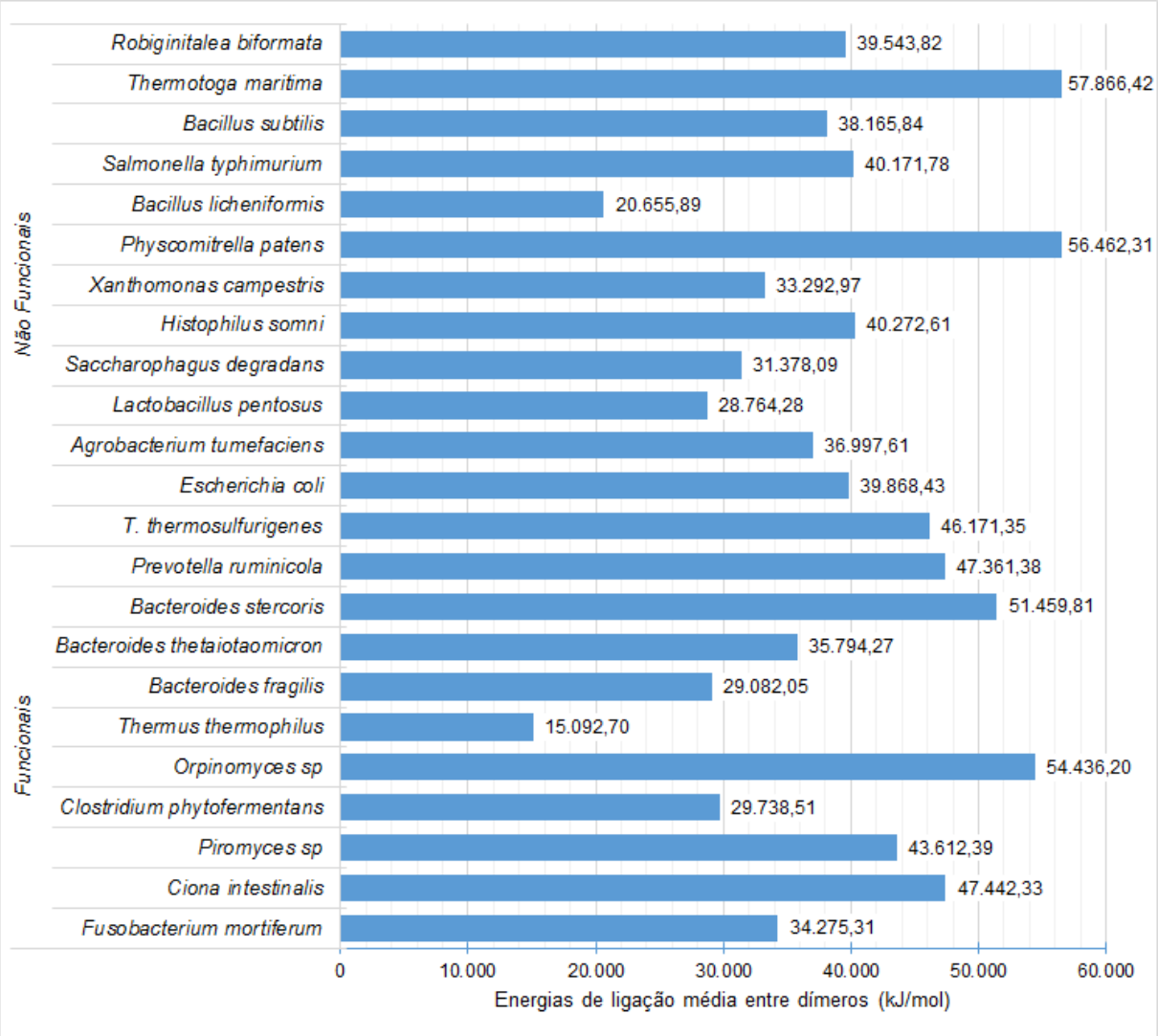


Figura 4.9: Gráfico das médias das energias de ligação entre os dímeros de cada xilose isomerase modelada após a minimização de energia.

Para comparar os dados, foram feitas distribuições dos valores de energia interna e de ligação, levando em conta qual dímero ou monômero era analisado e a qual grupo a proteína pertencia, a seguir são exibidos os gráficos que mostram essa distribuição.

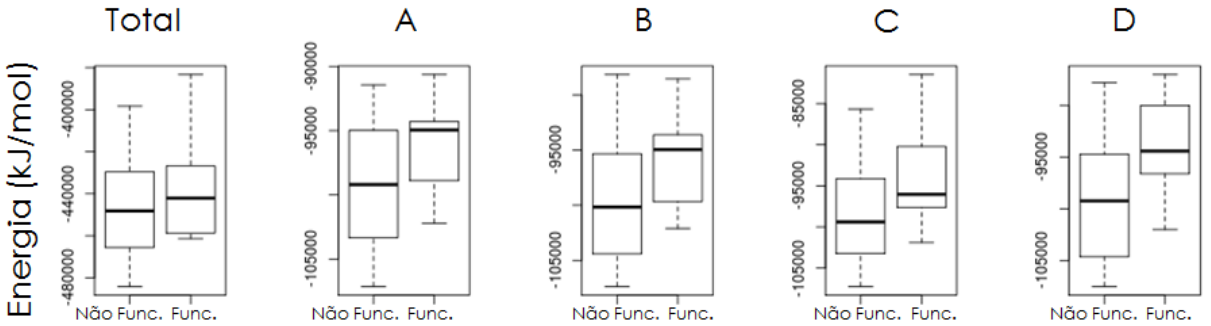


Figura 4.10: Distribuição estatística da energia interna total e de cada monômero entre as proteínas não funcionais e entre as proteínas funcionais.

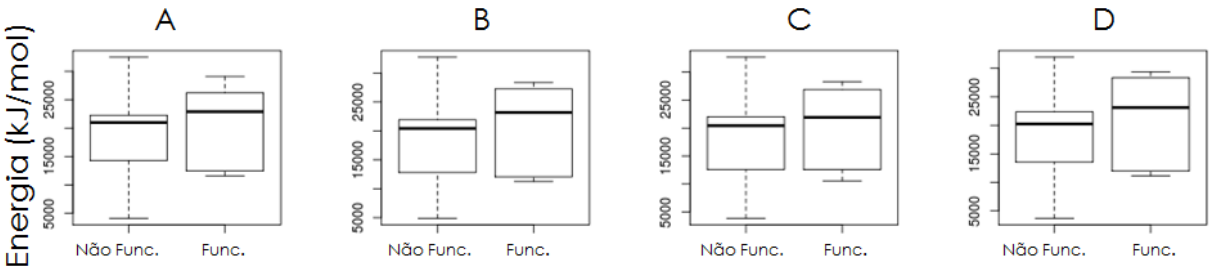


Figura 4.11: Distribuição estatística da energia de ligação de cada monômero com o restante da proteína entre as proteínas não funcionais e entre as proteínas funcionais.

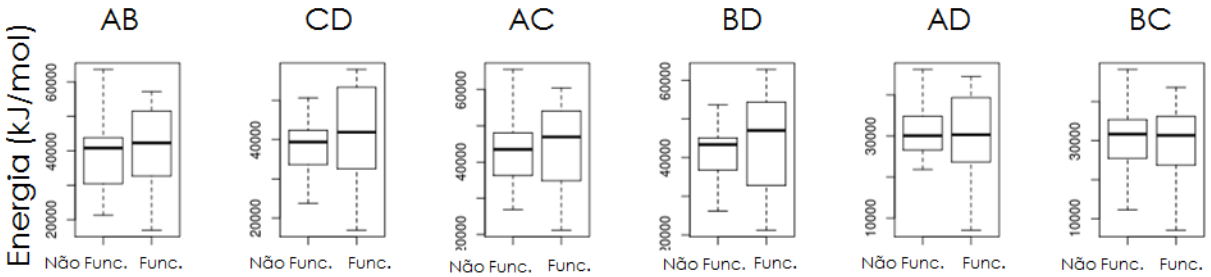


Figura 4.12: Distribuição estatística da energia de ligação de cada dímero com o restante da proteína (seu dímero complementar) entre as proteínas não funcionais e entre as proteínas funcionais.

Infelizmente, a análise dos gráficos e dos dados de energia não permitiu chegar a nenhuma conclusão ou indício de relação entre as energias internas ou de ligação com a funcionalidade das proteínas em *S. cerevisiae.* É possível notar, nas distribuições de energia, que as proteínas funcionais tendem a apresentar energias internas maiores (menos negativas) do que as não funcionais e uma energia de ligação maior, porém as diferenças não são tão definidas e o espaço amostral pequeno para ser possível tirar alguma conclusão.

A fim de averiguar as posições dos motivos na xilose isomerase e as possíveis relações entre eles, o modelo tridimensional da *Piromyces sp.*, escolhida por ter um

bom desempenho na *S. cerevisiae*, teve sua superfície colorida e seus motivos destacados. As figuras a seguir apresentam a estrutura tridimensional de vários ângulos e com alguns monômeros eventualmente omitidos, de modo a permitir a observação todas as relações entre os motivos.

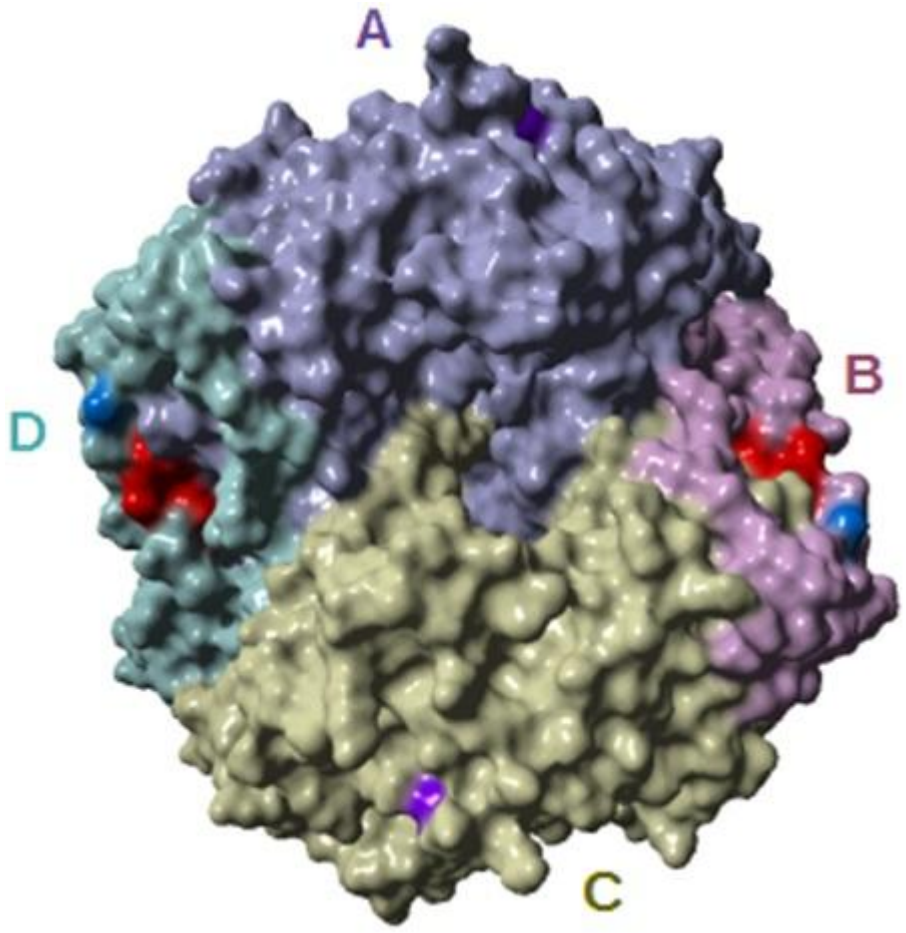


Figura 4.13: Posição dos monômeros no modelo tridimensional da xilose isomerase de *Piromyces sp.*

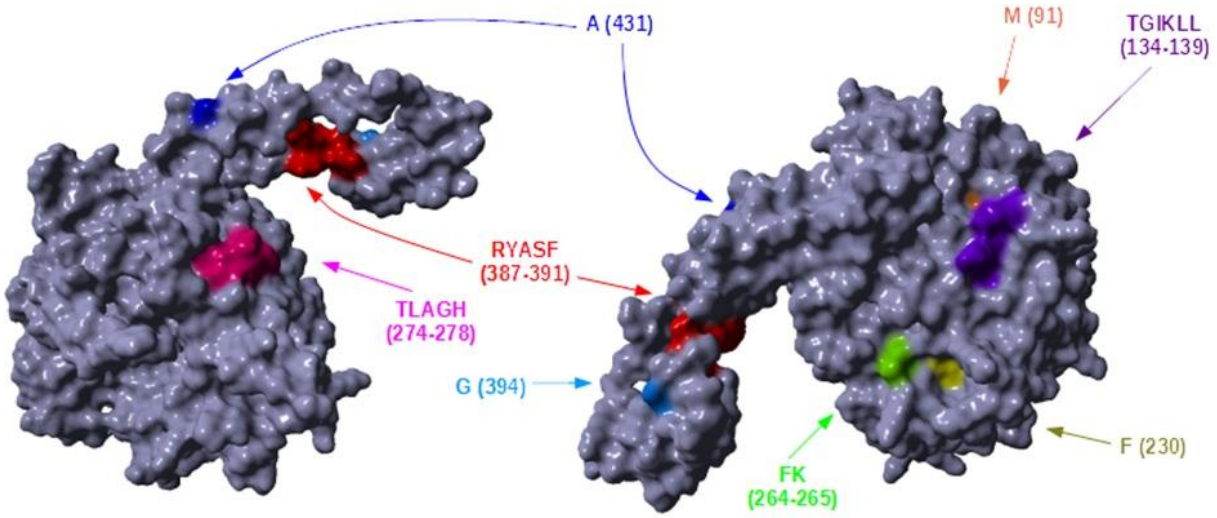


Figura 4.14: Motivos identificados em um monômero do modelo.

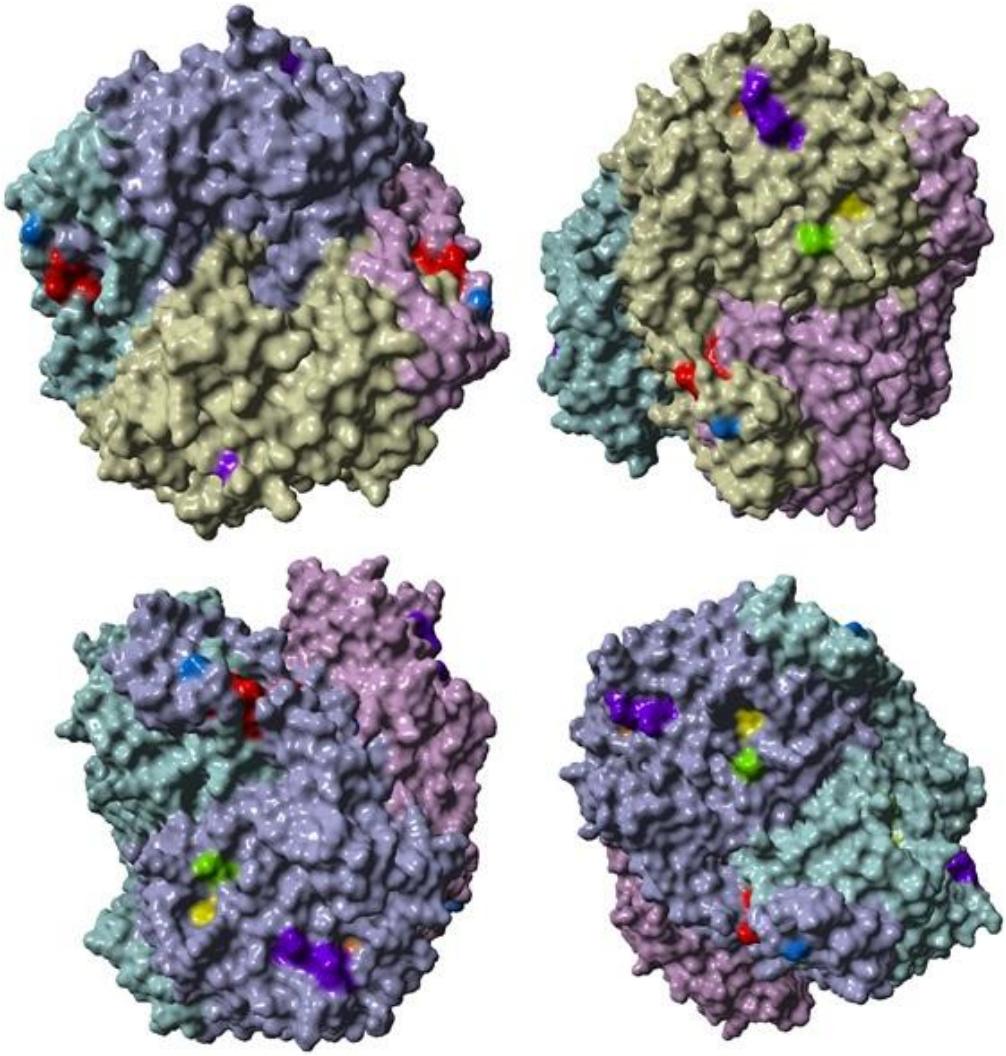


Figura 4.15: Modelo tridimensional da xilose isomerase de *Piromyces sp.* visto de vários ângulos

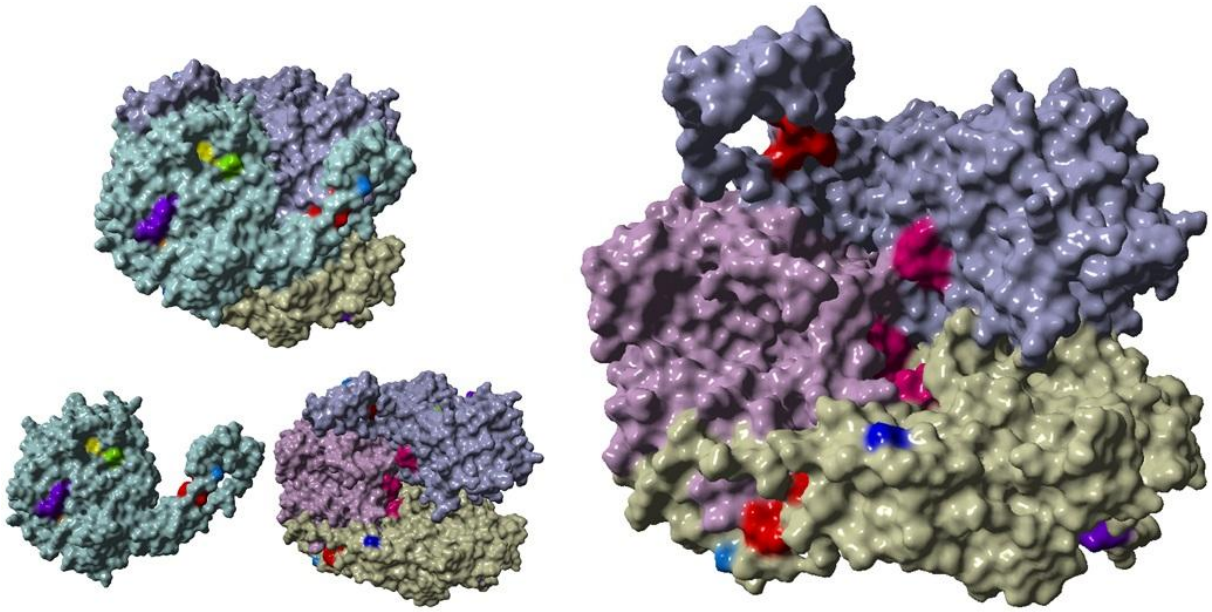


Figura 4.16: Vista do interior do modelo tridimensional ao ser retirado um monômero.

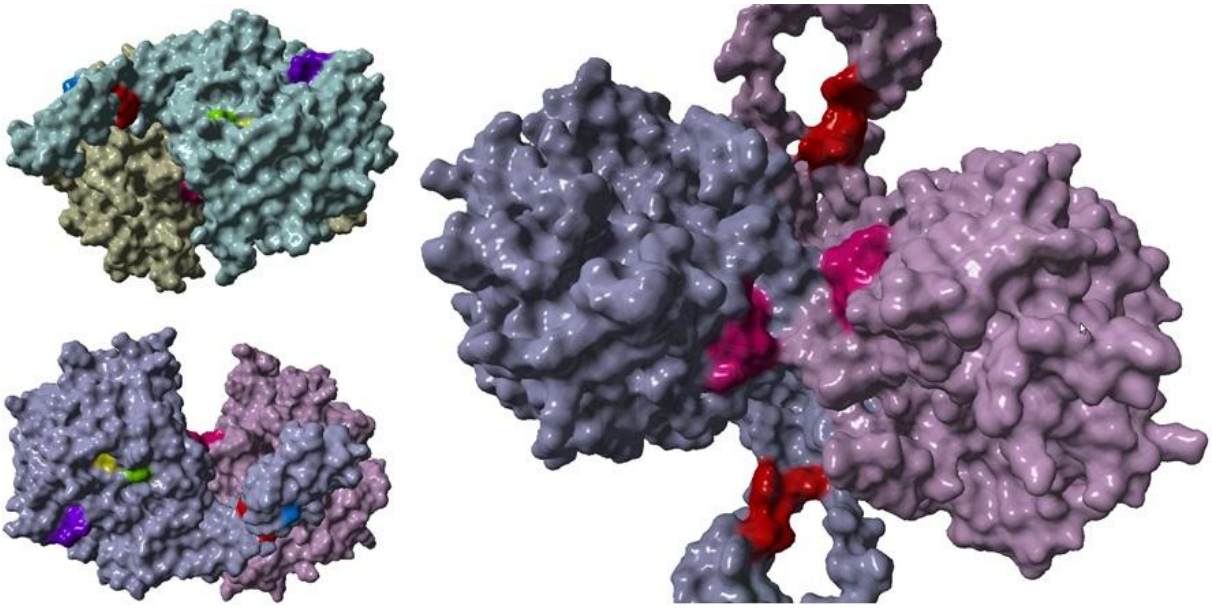


Figura 4.17: Vista do interior do modelo tridimensional ao ser retirado um dímero.

A observação da estrutura tridimensional revela que os motivos TGIKLL na posição 134, Metionina na posição 91, Fenilalanina na posição 230 e Fenilalanina seguida de Lisina na posição 264 se encontram na interface externa do tetrâmero, podendo representar sítios ativos, talvez essenciais para a conversão de xilose em xilulose; já o motivo TLAGH na posição 274 está totalmente na interface interna da proteína, exatamente na região onde os 4 monômeros fazem contato, sendo provavelmente essencial para dar estabilidade ao tetrâmero; por fim, os motivos

RYASF na posição 387, Glicina na posição 394 e Alanina na posição 431, além de também estarem na superfície externa, participam da interface entre monômeros, não sendo possível averiguar a priori que função desempenham.

Apesar das evidências de contribuição de cada motivo encontradas pela modelagem por homologia, outros estudos futuros devem ser realizados para confirmar as suspeitas, como estudos experimentais da proteína, ou, como alternativa menos custosa, mais estudos sobre os modelos tridimensionais criados. Por exemplo, a troca dos motivos por alanina e recálculo das energias internas e de ligação, a fim de verificar o quão importante são os motivos para a estabilidade da proteína.

Outra possiblidade é a modelagem de proteínas selecionadas pelo pipeline descrito na primeira parte do capítulo para posterior verificação de seus motivos, sendo que as xilose isomerases que apresentarem motivos em posições semelhantes às encontradas nas xilose isomerases funcionais, apresentam uma chance maior de também serem funcionais.

Capítulo 5

CONCLUSÃO

Este trabalho consistiu na utilização de ferramentas computacionais e biologia de sistemas para a busca de alternativas que permitam e melhorem a fermentação de xilose pela levedura *Saccharomyces cerevisiae* para produção de bioetanol de segunda geração. Na primeira etapa do trabalho, análises de balanço de fluxo (FBA) sobre variações de um modelo metabólico em escala genômica da levedura indicaram caminhos para a obtenção do balanço redox da via oxi-redutiva de metabolização de xilose; ora com o auxílio da via da fosfoquetolase (como já discutido em outros estudos), ora com auxílio das reações NADH quinase e NADP fosfatase.

Na segunda etapa, o pipeline desenvolvido permitiu identificação de xilose isomerases de outros organismos com potencial de apresentar funcionalidade em *S. cerevisiae*, facilitando as decisões de futuros trabalhos de bancada. Além disso esta etapa promoveu a modelagem por homologia de algumas xilose isomerases e posterior análise das energias e inspeção visual dos motivos presentes na proteína proveniente de *Piromyces sp.*

Além dos resultados obtidos, o presente estudo forneceu novos conhecimentos ao laboratório, permitindo que trabalhos futuros envolvendo outras proteínas e espécies possam ser realizados com mais facilidade, utilizando as técnicas, scripts e pipeline desenvolvidos neste projeto.

REFERÊNCIAS BIBLIOGRÁFICAS

AGBOGBO, F. K. et al. Fermentation of glucose/xylose mixtures using *Pichia stipites*. **Process Biochemistry**, Amsterdam v. 41, n. 11 p. 2333-2336, 2006.

AMORE, R. et al. Biotechnology the fermentation of xylose-an analysis of the expression of *Bacillus* and *Actinoplanes* xylose isomerase genes in yeast. **Applied microbiology and biotechnology**, v. 75, p.351-357, 1989.

ARGUESO, J. L. et al. Genome structure of a *Saccharomyces cerevisiae* strain widely used in bioethanol production. **Genome Research**, Cold Spring Harbor, v. 19, n. 12, p. 2258-2270, 2009. Disponível em <http://genome.cshlp.org/content/19/12/2258>. Acesso em: 4 fev. 2016.

AUNG, H. W. et al. Revising the representation of fatty acid, glycerolipid, and glycerophospholipid metabolism in the consensus model of yeast metabolism. **Industrial Biotechnology**, v. 9 n. 4, p. 215-228. Disponível em <http://online.liebertpub.com/doi/abs/10.1089/ind.2013.0013>. Acesso em: 4 fev. 2016.

BALAT, M. Production of bioethanol from lignocellulosic materials via the biochemical pathway: A review. **Energy Conversion and Management**, Amsterdam, v. 52 n. 2, p. 858–875, 2011.

BRAT, D. et al. Functional expression of a bacterial xylose isomerase in *Saccharomyces cerevisiae*. **Applied and environmental microbiology**, v. 75, n. 8, p.2304–2311, 2009.

BRO, C. et al. In silico aided metabolic engineering of *Saccharomyces cerevisiae* for improved bioethanol production. **Metabolic Engineering**, v. 8, n. 2, p. 102–111, 2006.

CAIMI, P. G. et al. E I Du Pont De Nemours And Company. **Pnp gene modification for improved xylose utilization in zymomonas**. *United States Patent Application Publication, patent number: US2013/ 0157331*, 2013

CAPRILES, P. V. et al. Modelos Tridimensionais. In: VERLI, H. (Org.). **Bioinformática:** da Biologia à Flexibilidade Molecular. São Paulo: SBBq, 2014. p. 148–171.

CARVALHO-NETTO, O. V. et al. *Saccharomyces cerevisiae* transcriptional reprograming due to bacterial contamination during industrial scale bioethanol production. **Microbial Cell Factories**, v.14, n. 13, 2015. Disponível em < http://microbialcellfactories.biomedcentral.com/articles/10.1186/s12934-015-0196-6>. Acesso em: 4 fev. 2016.

CHANDEL, A. K. et al. Detoxification of sugarcane bagasse hydrolysate improves ethanol production by *Candida shehatae* NCIM 3501. **Bioresource Technology**, Amsterdam, v. 98, n. 10, p. 1947–1950, 2007.

CHENG, K.K. et al. Sugarcane bagasse hemicelulose hydrolysate for ethanol production by acid recovery process. **Biochemical Engineering Journal**, Amsterdam, v.38, n. 1, p.105-109, 2008.

CORDIER, H. et al. A metabolic and genomic study of engineered *Saccharomyces cerevisiae* strains for high glycerol production. Metabolic Engineering, v. 9, n. 4, p. 364–378, 2007.

DOBSON; KRUCKEBERG, 2013. Butamax(Tm) Advanced Biofuels Llc. **Lignocellulosic hydrolysates as feedstocks for isobutanol fermentation**. *United States Patent Application Publication, patent number: US2013/0035515*, 2013.

EDDY, S. R. Profile hidden Markov models. **Bioinformatics**, v. 14 n. 9, p; 755-763, 1998.

GÁRDONYI, M.; HAHN-HÄGERDAL, B. The *Streptomyces rubiginosus* xylose isomerase is misfolded when expressed in *Saccharomyces cerevisiae*. **Enzyme and Microbial Technology**, v. 32, n. 2, p. 252–259, 2003.

GE, J. et al. Danisco Us Inc. **Method of using alpha-amylase from aspergillus clavatus for saccharification**. *United States Patent Application Publication, patent number: US2013/ 0323798*, 2013

GIBAS, C.; JAMBECK, P. Desenvolvendo Bioinformática. **Campus**, Rio de Janeiro, 2001

GÍRIO, F. M. et al. Hemicelluloses for fuel ethanol: A review. **Bioresource Technology**, Amsterdam, v. 101, n. 13, p. 4775–4800, 2010.

GOLDEMBERG, J. The Brazilian biofuels industry. **Biotechnology for Biofuels**, Londres, v.1, n. 6, 2008. Disponível em < http://biotechnologyforbiofuels.biomedcentral.com/articles/10.1186/1754-6834-1-6>. Acesso em: 4 fev. 2016.

HA, S. et al. Engineered *Saccharomyces cerevisiae* capable of simultaneous cellobiose and xylose fermentation. **PNAS**, v. 108, n. 2, p.1–6, 2010.

HEAVNER, B. L. et al. Yeast 5 – an expanded reconstruction of the *Saccharomyces cerevisiae* metabolic network. **BMC Systems Biology**, v. 6 n. 55, 2012. Disponível em <http://bmcsystbiol.biomedcentral.com/articles/10.1186/1752-0509-6-55>. Acesso em: 4 fev. 2016.

HEAVNER, B. L. et al. Version 6 of the consensus yeast metabolic network refines biochemical coverage and improves model performance. **Database: The Journal of Biological Databases and Curation**, bat059, 2013. Disponível em < http://database.oxfordjournals.org/content/2013/bat059>. Acesso em: 4 fev. 2016.

HECTOR, R.E. et al. Growth and fermentation of D-xylose by *Saccharomyces cerevisiae* expressing a novel D-xylose isomerase originating from the bacterium *Prevotella ruminicola* TC2-24. **Biotechnology for biofuels**, v. 6 n. 1, p. 84, 2013.

HERRGÅRD, M. J. et al. A consensus yeast metabolic network reconstruction obtained from a community approach to systems biology. **Nature Biotechnology**, v. 26 n. 10, p. 1155-1160, 2008. Disponível em < http://www.nature.com/nbt/journal/v26/n10/full/nbt1492.html>. Acesso em: 4 fev. 2016.

HITZ, W. D. et al. E I Du Pont De Nemours And Company. **Gene inactivation allowing immediate growth on xylose medium by engineered zymomonas**. *United States Patent Application Publication, patent number: US2013/ 0157332*, 2013

HOU, J. et al. Impact of overexpressing NADH kinase on glucose and xylose metabolism in recombinant xylose-utilizing *Saccharomyces cerevisiae*. **Applied Microbiology and Biotechnology**, v. 82, n. 5, p. 909-919, 2009.

HUGHES, S. R.; BUTT, T. R. **Transformed Saccharomyces cerevisiae Engineered for Xylose Utilization.** *United States Patent Application Publication, patent number: US2010/ 0112658,* 2010.

JORDAN, s. et al. Archer Daniels Midland Company. **Xylose isomerase and xylitol dehydrogenase combination for xylose fermentation to ethanol and *B. fragilis* xylose isomerase**. *United States Patent Application Publication, patent number: US2014/0017768*, 2014

KARHUMA, K. et al. Comparison of the xylose reductase-xylitol dehydrogenase and the xylose isomerase pathways for xylose fermentation by recombinant *Saccharomyces cerevisiae*. **Microbial Cell Factories**, Londres, v. 6, n. 5, 2007. Disponível em <http://microbialcellfactories.biomedcentral.com/articles/10.1186/1475-2859-6-5>. Acesso em: 4 fev. 2016.

KAWAI, S.; MURATA, K. Structure and Function of NAD Kinase and NADP Phosphatase: Key Enzymes That Regulate the Intracellular Balance of NAD(H) and NADP(H). **Bioscience, Biotechnology, and Biochemistry**, v. 72, n. 4, p. 919-930, 2008. Disponível em <http://www.tandfonline.com/doi/abs/10.1271/bbb.70738>. Acesso em: 4 fev. 2016.

KONAGURTHU, A. S. et al. MUSTANG: A multiple structural alignment algorithm. **Proteins**, v. 64, n. 3, p. 559–574, 2006.

KOOTSTRA, A. M. J. et al. Optimization of the dilute maleic acid pretreatment of wheat straw. **Biotechnology for Biofuels**, v. 2, n. 31, 2009. Disponível em < http://biotechnologyforbiofuels.biomedcentral.com/articles/10.1186/1754-6834-2-31>. Acesso em: 4 fev. 2016.

KRIEGER, E. et al. Making optimal use of empirical energy functions: Force-field parameterization in crystal space. **Proteins**, v. 57, n. 4, p. 678–683, 2004.

KRIEGER, E. et al. Homology Modeling. In: BOURNE P. E.; WEISSIG, H. (Org.). **Structural Bioinformatics**, Hoboken : John Wiley & Sons, 2005. v. 44, p. 507–521. Disponível em <http://www.cmbi.ru.nl/edu/bioinf4/articles/homologymodeling.pdf>. Acesso em: 4 fev. 2016.

KUEPFER, L. et al. Metabolic functions of duplicate genes in *Saccharomyces cerevisiae*. **Genome Research**, Cold Spring Harbor, v. 15, p. 1421-1430, 2005. Disponível em < http://genome.cshlp.org/content/15/10/1421>. Acesso em: 04 fev. 2016.

KUYPER, M. et al. Minimal metabolic engineering of *Saccharomyces cerevisiae* for efficient anaerobic xylose fermentation: a proof of principle. **FEMS Yeast Research**, Delft, v. 4, n. 6, p. 655–664, 2004. Disponível em <http://onlinelibrary.wiley.com/doi/10.1016/j.femsyr.2004.01.003/full>. Acesso em: 04 fev. 2016.

KUYPER, M. et al. Metabolic engineering of a xylose-isomerase-expressing *Saccharomyces cerevisiae* strain for rapid anaerobic xylose fermentation. **FEMS Yeast Research**, Delft, v. 5, n. 4, p. 399–409, 2005. Disponível em <http://onlinelibrary.wiley.com/doi/10.1016/j.femsyr.2004.09.010/full>. Acesso em: 04 fev. 2016.

KUYPER, M. et al. High-level functional expression of a fungal xylose isomerase: the key to efficient ethanolic fermentation of xylose by yeasts? **FEMS Yeast Research**, v. 4, n. 1, p. 69–78, 2003.

LAKSHMANAN, M. Software applications for flux balance analysis. **Briefings in Bioinformatics**, v. 15 n. 1, p. 108-122, 2012. Disponível em <http://bib.oxfordjournals.org/content/15/1/108.full>. Acesso em: 04 fev. 2016.

LARKIN, M. A. Clustal W and Clustal X version 2.0. **Bioinformatics**, v. 23 n. 21, p; 2947-2948, 2007.

LESK, A. M. Introduction to Bioinformatics. **Oxford University Press**, Oxford, 2006.

LIU, L. et al. Use of genome-scale metabolic models for understanding microbial physiology. **FEBS Letters**, v. 584, n. 12, p. 2556–2564, 2010. Disponível em <http://onlinelibrary.wiley.com/doi/10.1016/j.febslet.2010.04.052/full>. Acesso em: 04 fev. 2016.

MADHAVAN, A. et al. Xylose isomerase from polycentric fungus *Orpinomyces*: gene sequencing, cloning, and expression in *Saccharomyces cerevisiae* for bioconversion of xylose to ethanol. **Applied microbiology and biotechnology**, v. 82, n. 6, pp.1067–1078, 2009

MARGARITIS, A.; BAJPAI, P. Direct Fermentation of D-Xylose to Ethanol by *Kluyveromyces marxianus* Strains. **Applied and Environmental Microbiology**, Washington, v. 44, n. 5, p. 1039-1041, 1982. Disponível em <http://aem.asm.org/content/44/5/1039.full.pdf>. Acesso em: 02 fev. 2016.

MARIS, A. J. et al. Alcoholic fermentation of carbon sources in biomass hydrolysates by *Saccharomyces cerevisiae*: current status. **Antonie van Leeuwenhoek**, v. 90, n. 4, p.391–418, 2006.

MO, M. L. et al. Connecting extracellular metabolomic measurements to intracellular flux states in yeast. **BMC Systems Biology**, v. 3, n. 37, 2009. Disponível em < http://bmcsystbiol.biomedcentral.com/articles/10.1186/1752-0509-3-37>. Acesso em: 04 fev. 2016.

MOES, C. J. et al. Cloning and expression of the *Clostridium thermosulfurogenes* D-xylose isomerase gene (*xylA*) in *Saccharomyces cerevisiae*. **Biotechnology letters**, v. 18 n. 3, p. 269-274, 1996.

NISSEN, T. L. et al. Optimization of Ethanol Production in Saccharomyces cerevisiae by Metabolic Engineering of the Ammonium Assimilation. **Metabolic Engineering**, v. 2, n. 1, p. 69–77, 2000.

OLIVEIRA, M. Entre açúcares e genes. **Pesquisa Fapesp**, ed. 200, p. 86-91, 2012. Disponível em <http://revistapesquisa.fapesp.br/2012/10/11/entre-acucares-e-genes>. Acesso em: 02 fev. 2016.

ORTH, J. D. et al. What is flux balance analysis? **Nature Biotechnology**, v. 28 n. 3 p. 245-248, 2010.

ÖSTERLUND, T. et al. Fifteen years of large scale metabolic modeling of yeast: Developments and impacts. **Biotechnology Advances**, v. 30, n. 5, p. 979–988, 2012.

OTERO J. M.; NIELSEN J. Industrial systems biology. **Biotechnology and Bioengineering**, v. 105, n. 3, p. 439–460, 2010.

OTERO R. R., et al. Forskarpatent I Syd Ab. **Xylose isomerase with improved properties**. *United States Patent Application Publication, patent number: US6475768*, 2002.

PARACHIN, N. S. et al. Kinetic modelling reveals current limitations in the production of ethanol from xylose by recombinant *Saccharomyces cerevisiae*. **Metabolic Engineering**, v. 13, n. 5, p. 508–517, 2011.

RENEWABLE FUELS ASSOCIATION – RFA. Accelerating industry innovation. **2012 Ethanol industry outlook**, Washington DC, p. 3-23, 2012. Disponível em: <http://www.ethanolrfa.org/wp-content/uploads/2015/09/2012-Ethanol-Industry-Outlook.pdf>. Acesso em: 02 fev. 2016.

RAJGARHIA R. R., et al. Cargill, Incorporated. **Genetically modified yeast species, and fermentation processes using genetically modified yeast**. *United States Patent Application Publication, patent number: US8440451*, 2013.

RUDOLF, A. et al. Ethanol production from traditional and emerging raw materials. In: SATYANARAYANA, T.; KUNZE, G. **Yeast Biotechnology:** Diversity and Applications. Heidelberg: Springer, 2009. cap. 23, p. 489-514.

SANNY, T.; STARK B. C. **Increased Ethanol Production By Genetic Engineering Of Microorganisms To Express Hemoglobin**. *United States Patent Application Publication, patent number: US2011/ 0269200*, 2011.

SANTOS FILHO, O. A.; ALENCASTRO, R. B. Modelagem de proteínas por homologia. **Química Nova**, São Paulo , v. 26, n. 2, p. 253-259, mar. 2003. Disponível em

<http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0100-40422003000200019>. Acesso em 04 fev. 2016.

SARTHY, A. V. et al., 1987. Expression of the *Escherichia coli* xylose isomerase gene in *Saccharomyces cerevisiae*. **Applied and environmental microbiology**, v. 53, n. 9, p.1996–2000, 1987.

SATOSHI et al. Riken, Kabushiki Kaisha Toyota Chuo Kenkyusho. **Xylose isomerase and use thereof**. *United States Patent Application Publication, patent number: US2013/0095538*, 2013

SAXENA, A. et al. Fundamentals of Homology Modeling Steps and Comparison among Important Bioinformatics Tools: An Overview. **Science International**, v. 1, n. 7, p. 237–252, 2013. Disponível em <http://www.scienceinternational.com/fulltext/?doi=sciintl.2013.237.252>. Acesso em 04 fev. 2016.

SEGRÈ, D. et al. Analysis of optimality in natural and perturbed metabolic networks. PNAS, v. 99 n. 23, p. 15112-15117, 2002. Disponível em <http://www.pnas.org/content/99/23/15112>. Acesso em 04 fev. 2016.

SCHELLENBERGER, J. et al. Quantitative prediction of cellular metabolism with constraint-based models: the COBRA Toolbox v2.0. **Nature Protocols**, v. 6, n. 9, p. 1290-1307, 2011. Disponível em <http://www.nature.com/nprot/journal/v6/n9/full/nprot.2011.308.html>. Acesso em 04 fev. 2016.

SHEN, Y. et al. An efficient xylose-fermenting recombinant *Saccharomyces cerevisiae* strain obtained through adaptive evolution and its global transcription profile. **Applied microbiology and biotechnology**, v. 96, n. 4, p.1079–91, 2012.

SONDEREGGER, M. et al. Metabolic Engineering of a Phosphoketolase Pathway for Pentose Catabolism in *Saccharomyces cerevisiae.* **Applied and Environmental Microbiology**, v. 70, n. 5, p. 2892-2897, 2004. Disponível em: <http://aem.asm.org/content/70/5/2892>. Acesso em: 04 fev. 2016.

SPAANS, S. K. et al. NADPH-generating systems in bacteria and archaea. Frontiers in Microbiology, v. 6, p. 742, 2015. Disponível em: <http://journal.frontiersin.org/article/10.3389/fmicb.2015.00742/full>. Acesso em: 04 fev. 2016.

SUBBIAN, E. et al. Codexis, Inc. **Pentose Fermentation By a Recombinant Microorganism**. *United States Patent Application Publication, patent number: US2011/0294170*, 2011

SUBBIAN, E. et al. Codexis, Inc. **Pentose Fermentation By a Recombinant Microorganism**. *United States Patent Application Publication, patent number: US2013/0004998*, 2013

TEUNISSEN, A. W.; DE BONT, J. A. **Xylose isomerase genes and their use in fermentation of pentose sugars**. *United States Patent Application Publication, patent number: US2011/0318790,* p.18, 2011.

WALFRIDSSON, M. et al. Ethanolic fermentation of xylose with *Saccharomyces cerevisiae* harboring the *Thermus thermophilus xylA* gene, which expresses an active xylose (glucose) isomerase. **Applied and environmental microbiology**, v. 62, n. 12, p.4648–4651, 1996.

WINKLER, A. A. et al. Technische Universiteit Delft. **Metabolic engineering of xylose-fermenting eukaryotic cells**. *United States Patent Application Publication, patent number: US2012/0225451,* 2012.

ANEXOS

UNICAMP
COORDENADORIA DE PÓS-GRADUAÇÃO
INSTITUTO DE BIOLOGIA
Universidade Estadual de Campinas
Caixa Postal 6109. 13083-970, Campinas, SP, Brasil
Fone (19) 3521-6378. email: cpgib@unicamp.br
IB

DECLARAÇÃO

Em observância ao **§5º do Artigo 1º da Informação CCPG-UNICAMP/001/15**, referente a Bioética e Biossegurança, declaro que o conteúdo de minha Dissertação de Mestrado, intitulada "***Aplicação de ferramentas de Biologia de Sistemas em levedura industrial para produção de bioetanol de segunda geração***", desenvolvida no Programa de Pós-Graduação em Genética e Biologia Molecular do Instituto de Biologia da Unicamp, não versa sobre pesquisa envolvendo seres humanos, animais ou temas afetos a Biossegurança.

Assinatura: _____
Nome do(a) aluno(a): Felipe Alonso Martins

Assinatura: _____
Nome do(a) orientador(a): Gonçalo Amarante Guimarães Pereira

Data: 16/12/2015

Profa. Dra. Rachel Meneguello
Presidente
Comissão Central de Pós-Graduação
Declaração

As cópias de artigos de minha autoria ou de minha co-autoria, já publicados ou submetidos para publicação em revistas científicas ou anais de congressos sujeitos a arbitragem, que constam da minha Dissertação/Tese de Mestrado/Doutorado, intitulada **Aplicação de ferramentas de Biologia de Sistemas em levedura industrial para produção de bioetanol de segunda geração**, não infringem os dispositivos da Lei n.° 9.610/98, nem o direito autoral de qualquer editora.

Campinas, 16/12/2015

Assinatura : _____
Nome do(a) autor(a): **Felipe Alonso Martins**
RG n.° 44.607.403-2

Assinatura : _____
Nome do(a) orientador(a): **Gonçalo Amarante Guimarães Pereira**
RG n.° 1.713.878 SSP/BA